Statistical Methods for the Fraction who Benefit using a Randomized Trial

by

Emily J. Huang

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

April, 2017

Abstract

I develop new statistical methods to learn the fraction who benefit from treatment, using randomized trial data. Defined in the potential outcomes framework, the fraction who benefit is the proportion of patients whose potential outcome under treatment is better than that under control. Since only one potential outcome can be observed for any given person, this fraction is a non-identifiable parameter unless strong assumptions are made about the joint probability distribution of the potential outcomes. A strength of my methods is that they do not require assumptions about the joint distribution, but can incorporate a wide range of user-defined assumptions based on subject matter knowledge. I derive sharp lower and upper bounds on the fraction and develop a method to estimate these bounds, which is consistent and computationally efficient. The method is illustrated through an application to the MISTIE II (Minimally Invasive Surgery for Intracerebral Hemorrhage Evacuation Phase II) randomized trial on intracerebral hemorrhage. In addition, I develop a method to construct a confidence interval for the fraction who benefit, and prove that it is pointwise consistent. The method is demonstrated through an application

ABSTRACT

to the CLEAR III (Clot Lysis: Evaluating Accelerated Resolution of Intraventricular Haemorrhage III) randomized trial on intraventricular hemorrhage.

Advisor: Michael Rosenblum

Committee: Caleb Alexander, Karen Bandeen-Roche, Daniel Hanley, Ravi Varadhan

Alternates: Matthew Eisenberg, Daniel Scharfstein

Acknowledgments

To my advisor Michael Rosenblum, thank you for your mentorship during the past five years. I have really enjoyed working with you. You have a constant energy and enthusiasm that makes me feel excited about doing statistics. You are both intuitive and rigorous at the same time; this helped me understand many concepts deeper. Thank you for the experience of developing new statistical methods. It was fun to address real-world questions with new, simple approaches. Of course, thank you for your time. You met with me weekly for five years. This opportunity to ask you questions and chat with you about my ideas was invaluable to my learning.

To my EBA advisor Ravi Varadhan, thank you for your mentorship, time, and patience. I feel very lucky to have worked with you for over three years because you are a deeply passionate and thoughtful researcher, as well as a friendly person. I have always enjoyed our meetings since we talk about statistics and life. I am grateful that you introduced me to survival analysis and aging research through our work on analyzing grouped survival data. Also, I appreciate that you often dig into my code and give me advice about how to make it better.

ACKNOWLEDGMENTS

To my collaborators Dan Hanley and Wendy Ziai, thank you for the opportunity to address stroke questions together. Through our projects and meetings, I learned a lot about stroke and the trials that you have conducted. I also enjoyed the challenge of analyzing real-world data. Dan, thank you for your support of my statistical method development for the fraction who benefit. Moreover, thank you for the opportunity to apply my methods to MISTIE II and CLEAR III, and for sharing your clinical expertise to inform my data analyses.

To Karen Bandeen-Roche, thank you for getting me to think harder about my research. In particular, some questions you asked during my preliminary oral exam helped me pinpoint critical things that I needed to work on. Also, thank you for meeting with me whenever I needed advice, especially before milestones in the PhD program. Thank you for your positive feedback on my EBA talks and 621 labs. Your encouragement helped me feel more confident. Also, thank you for telling me about the lead TA position of the 620's during my second year. I am grateful that you made this gesture. Being a lead TA has been one of the best parts of my PhD experience.

To Caleb Alexander, thank you for your guidance on my fraction who benefit work. You helped me think about identifying new data sets that I can apply my methods to and how to share my work with interested parties. In addition, thank you for telling me about the CERSI program and encouraging me to apply for it. I was thrilled to be selected as a CERSI Scholar. The CERSI program has introduced me to people I would otherwise never have met and given me opportunities to present

my research at major conferences.

To my collaborator Ethan Fang, I have enjoyed working with you since you are friendly, enthusiastic, and energetic. Thank you for being a co-author on my papers. I am extremely grateful for your help with the proofs about linear and quadratic programming. Also, thank you for your comments on the writing and content of my papers.

In addition to research, a major part of my PhD experience was teaching. I would like to thank Margaret Taub, Scott Zeger, and Leah Jager for my first teaching experience as a TA for the undergraduate biostatistics course. I am especially grateful to Margaret for doing recitations with me. She helped me get comfortable with talking in front of an audience and answering questions on the spot. I would like to thank Karen Bandeen-Roche and Marie Diener-West for the opportunity to be a lead teaching assistant for 621 for three years. This experience gave me intense joy and satisfaction, and made me realize that I love teaching. Thank you Karen and Marie for being great role models. In preparation for my labs, I listened to many of your lectures and always felt you were caring of and patient with your students. Moreover, thank you Marie for responding to my many emails about the tough questions I got in lab.

To my parents and sister Jenny, thank you for your continuous support during my classes, qualifying exams, and dissertation research. It was a challenge for me to transition from engineering to statistics. I sometimes doubted whether I could get

through, but you helped me believe I could do it. I believe I would not have completed this journey without your help.

To my friends, thank you for making my social life a lot of fun. I enjoyed hanging out, trying our restaurants, and playing games with you. Also, thank you for being willing to chat with me whenever I was struggling. In particular, David Lenis, Helen Powell, Elizabeth Sweeney, Therri Usher, and Nicole Armstrong were very supportive of me, and I am extremely grateful to them.

I was lucky to be supported by the Epidemiology and Biostatistics of Aging (EBA) training grant for five years. I loved being part of the EBA community, which is a fun and friendly group that is passionate about aging research. I also enjoyed giving and hearing talks about aging during EBA meetings.

Contents

Abstract **ii**

Acknowledgments **iv**

List of Tables **x**

List of Figures **xi**

1 Introduction **1**

2 Estimating Bounds on the Fraction who Benefit using Randomized Trial Data **4**

2.1 Introduction 4

2.2 MISTIE II Trial 8

2.3 Bound Parameters 9

2.4 Bound Estimators 23

2.5 MISTIE Application using Bound Estimator from Section 2.4.2 . . . 49

2.6 Simulation Studies 60

2.7 Discussion 68

2.8 Software 69

3 Constructing a Confidence Interval for the Fraction who Benefit Using a Randomized Trial 71

3.1 Abstract 71

3.2 Introduction 72

3.3 Related Work 76

3.4 Notation, Parameter Definition, and Assumptions 78

3.5 Proposed Method 82

3.6 Simulation Studies 103

3.7 Application to the CLEAR III trial 110

3.8 Discussion 115

4 Discussion 117

5 Appendix 119

List of Tables

2.1 Bias and Standard Error of Proposed Estimator. 62
2.2 Coverage Probabilities and Average Widths of CI's Constructed using n-bootstrap and m-out-of-n bootstrap. 64
2.3 RICV5: Estimator Properties as a Function of Number of Strata. . . 67
2.4 180-day mRS Score: Estimator Properties as a Function of Number of Strata. 67

3.1 Simulation Settings. 105
3.2 95% CI's for the Fraction who Benefit from Alteplase Compared to Saline. 113
3.3 Sample Sizes for 95% CI's in Table 3.2. 114

5.1 Estimated Bounds on the Fraction who Benefit, with respect to 30-day mRS Score. 119
5.2 Estimated Bounds on the Fraction who Benefit, with respect to 180-day mRS Score. 120
5.3 Estimated Bounds on the Fraction who Benefit, with respect to RICV5. 120
5.4 Estimated Bounds on the Fraction who Benefit, with respect to RICV2. 121
5.5 Estimated Bounds on the Fraction who Benefit, with respect to RICV1. 121
5.6 Estimated Bounds on the Fraction who Benefit, with respect to RICV2. 122

List of Figures

2.1 Joint Distribution of the Potential Outcomes. 14
2.2 Support Restriction: Benefit ≤ 3 levels. 14
2.3 Distributions of $\sqrt{n}(\overline{\psi}_l - \psi_l(P_0))$ under P_0 (left panel), and of $\sqrt{n}(\overline{\psi}_l - \psi_l(P^{(n)}))$ under $P^{(n)}$ (right panel), for $n = 100, 1000, 10000$, and 1000000. 40
2.4 Empirical Distribution of 30-day mRS Score under Treatment and Control. 51
2.5 Empirical Distribution of 180-day mRS Score under Treatment and Control. 52
2.6 Bounds Estimates for 30-day mRS Score, using Method from Section 2.4.2. 53
2.7 Bounds Estimates for 180-day mRS Score, using Method from Section 2.4.2. 53
2.8 Empirical Probability Mass Functions of RICV5 under Treatment and Control. 59
2.9 Bound Estimates for RICV5, using Method from Section 2.4.2. 60

3.1 Coverage probabilities in Setting A at $n = 500$. *The grey region spans from $\psi = 0$ to $\psi = 0.5$, which are the lower and upper bounds $\psi_l^{\mathcal{R}}$ and $\psi_u^{\mathcal{R}}$ in Setting A. To achieve good coverage under Assumption 3, coverage probabilities should be ≥ 0.95 for all ψ in the grey region. For legibility of the plot, the curves for $m = 0.9n$ and $m = 0.75n$ are not shown. They lie between the curves for $m = n$ and $m = 0.5n$, but closely resemble the curve for $m = n$.* 108
3.2 Coverage probabilities in Setting B at $n = 500$. *The grey region is the single point $\psi = 0$, since in Setting B the lower and upper bounds $\psi_l^{\mathcal{R}}$ and $\psi_u^{\mathcal{R}}$ are both zero. To achieve good coverage under Assumption 3, coverage probabilities should be ≥ 0.95 at $\psi = 0$. For legibility of the plot, the curves for $m = 0.9n$ and $m = 0.75n$ are not shown. They lie between the curves for $m = n$ and $m = 0.5n$, but closely resemble the curve for $m = n$.* 109

LIST OF FIGURES

5.1 Empirical Probability Mass Functions of RICV2 under Treatment and Control. 123
5.2 Empirical Probability Mass Functions of RICV10 under Treatment and Control. 124
5.3 Empirical Probability Mass Functions of RICV20 under Treatment and Control. 125

Chapter 1

Introduction

In many randomized controlled trials, the primary analysis focuses on the average treatment effect and does not address whether treatment benefits are widespread or limited to a select few. This problem affects many disease areas, since it stems from how randomized trials, often the gold standard for evaluating treatments, are designed and analyzed. My aim is to make inferences about the fraction who benefit from treatment, using randomized trial data. This fraction is formally defined as the proportion of patients whose potential outcome under treatment is better than that under control. In other words, it is the fraction who are better off with treatment. The fraction who benefit is of interest to patients and doctors selecting between treatment and control. In addition, it can be a useful tool for medical researchers. For example, if the fraction is small, this is a signal to researchers that they should devote resources toward identifying the exclusive subgroup that benefits from treatment.

Since we can observe only one potential outcome per person (Holland, 1986), the fraction who benefit is typically non-identifiable without strong, untestable assumptions about the joint distribution of the potential outcomes. Examples of such assumptions include independence of the potential outcomes, or conditional independence of the potential outcomes given baseline covariates collected in the trial (Zhang et al., 2013). To maximize the applicability of my methods, I avoid requiring strong assumptions. The two methods proposed in Chapters 2 and 3 of this thesis do not require any assumptions about the joint distribution of the potential outcomes. However, they allow the user to incorporate certain assumptions based on subject matter knowledge. The user-defined assumptions that can be incorporated are restrictions on the support of the joint distribution of the potential outcomes. One example is the no harm assumption, which is that for any given person the potential outcome under treatment is no worse than the potential outcome under control.

In Chapter 2, I derive and estimate sharp lower and upper bounds on the fraction who benefit. The main contributions in this chapter include (i) deriving bound parameters that are identifiable, (ii) proving the plug-in estimator can be inconsistent if support restrictions are made on the joint distribution of the potential outcomes; (iii) developing the first consistent estimator for this case; (iv) applying this estimator to the MISTIE II (Minimally Invasive Surgery for Intracerebral Hemorrhage Evacuation Phase II) randomized trial to determine whether the estimates can be informative. The proposed estimator is computed using linear programming, which has readily

available software that is efficient. The bound estimates can typically be computed in under a second.

In Chapter 3, I address the question of how to construct a confidence interval for the fraction who benefit. The main contributions in this chapter include (i) developing a new method to construct a confidence interval for the fraction using randomized trial data, (ii) proving the confidence interval is pointwise consistent, (iii) evaluating its empirical coverage and width at finite sample sizes through simulation, (iv) applying it to the CLEAR III (Clot Lysis: Evaluating Accelerated Resolution of Intraventricular Haemorrhage III) randomized trial to compare alteplase versus saline in treating patients with severe stroke. In simulation, the proposed method can have substantially smaller average width compared to an existing method using the m-out-of-n bootstrap on the bounds. Also, the new method is computationally efficient because it utilizes quadratic programming.

The methods in Chapters 2 and 3 are designed for an ordinal outcome, with binary outcomes as a special case. In Chapter 4, I recapitulate the main contributions of my thesis work and discuss future directions.

Excerpts of this chapter and Chapter 2 are from the paper entitled "Inequality in Treatment Benefits: Can We Determine if a New Treatment Benefits the Many or the Few?", which has been published in *Biostatistics* (Huang et al., 2017).

Chapter 2

Estimating Bounds on the Fraction who Benefit using Randomized Trial Data

2.1 Introduction

We aim to estimate bounds on the fraction of the population who benefit from a treatment. This fraction is defined in the potential outcomes framework. Each participant has two potential outcomes, one representing the participant's outcome if assigned to treatment and the other if assigned to control. The fraction who benefit is defined as the fraction of the population whose potential outcome under treatment is better than that under control. In other words, it is the fraction who would be

better off under treatment. Without strong assumptions, the fraction who benefit is generally a non-identifiable parameter. This is because only one potential outcome can be observed for any given individual (Holland, 1986). However, sharp bounds on the fraction are identifiable (Williamson and Downs, 1990; Manski, 1997; Fan and Park, 2009, 2010; Kim, 2014).

We derive bounds using the marginal distributions of the potential outcomes, and show these bounds can be narrowed using a prognostic baseline variable and/or user-defined assumptions. The type of user-defined assumptions considered are restrictions on the support of the joint distribution of the potential outcomes, e.g., the no harm assumption. Our main contributions include (i) proving the plug-in estimator of the bounds can be inconsistent if support restrictions are made; (ii) developing the first consistent estimator for this case; (iii) applying this estimator to a randomized trial data set of a medical treatment to determine whether the estimates can be informative. We assume a simple randomized trial design, i.e., each participant's treatment assignment is an independent draw from a Bernoulli distribution. Our estimator can be computed using linear programming, i.e., the optimization of a linear objective function subject to linear equality and inequality constraints (Vanderbei, 2013). The bound estimates are typically computed in under a second.

We apply our estimator to the MISTIE II (Minimally Invasive Surgery for Intracerebral Hemorrhage Evacuation Phase II) randomized trial (Morgan et al., 2008; Hanley et al., 2016), which compared a new surgical intervention for stroke to stan-

dard medical management. As an example of our results in one case, the lower and upper bound estimates on the fraction who benefit are 0.10 and 0.73 when the outcome is a rating of functional disability 180 days post-stroke, and 0.82 and 0.96 when the outcome is reduction in clot volume.

Related work includes Manski (1997); Gadbury et al. (2004); Fan and Park (2009, 2010); Zhang et al. (2013); Kim (2014); Borusyak (2015). Fan and Park (2009, 2010) prove sharp bounds on the fraction who benefit, given the marginal distributions of the potential outcomes. Kim (2014) tightens those bounds using support restrictions on the joint distribution. Both propose estimators for their respective bounds. A key difference of our work, compared to Fan and Park (2009, 2010) and Kim (2014), is that we handle an ordinal outcome, while they handle a continuous outcome. Applying their formulae to an ordinal outcome can yield erroneous results (Section 2.3.1). Unlike Fan and Park (2009, 2010), we allow the incorporation of support restrictions, which leads to a more challenging estimation problem. We propose a new estimator that can be computed efficiently using linear programming. In contrast, the estimator of Kim (2014) generally requires solving a non-convex optimization problem to incorporate support restrictions. Non-convex problems are much more computationally difficult than linear programs.

Gadbury et al. (2004) derive bounds on the fraction who are harmed, given the marginal distributions of the potential outcomes, when the outcome is binary and no baseline variable/support restrictions are used (see Section 2.3.1); in contrast, we

consider ordinal outcomes, baseline variables, and support restrictions. For an ordinal outcome, Borusyak (2015) shows that sharp bounds on the fraction who benefit can be computed with linear programming. Borusyak does not address estimation of these bounds, which is a focus of our work.

Manski (1997) derives sharp bounds on the fraction who benefit given the no harm assumption, without using the marginal distributions of the potential outcomes. We can incorporate the no harm assumption but do not require it. We use the marginal distributions since they are identifiable in the randomized trial context. The method by Manski is designed for settings where treatment assignment is not random. Zhang et al. (2013) estimate the fraction who benefit, rather than bounds on it, by assuming the potential outcomes are conditionally independent given a set of baseline covariates collected in the trial.

The structure of this chapter is as follows. The MISTIE II trial is introduced in Section 2.2. The bound parameters are defined in Section 2.3. In Section 2.4, we prove that the plug-in estimator can be inconsistent, propose a new estimator that is consistent and computationally efficient, and discuss inference based on this estimator (which is challenging due to potential non-regularity). We apply our estimator to MISTIE II in Section 2.5, and present simulation results in Section 2.6. Future directions are discussed in Section 2.7.

2.2 MISTIE II Trial

MISTIE II is a recently completed Phase II randomized trial for intracerebral hemorrhage (ICH), a type of stroke that can impair cognitive/motor functions and cause death (Morgan et al., 2008; Hanley et al., 2016). The MISTIE II trial assessed the effectiveness of image-guided minimally invasive surgery (i.e., treatment), relative to standard medical care (i.e., control). There were 96 participants, with 54 assigned to treatment and 42 to control. The randomization ratio gave a higher likelihood of being assigned to treatment, yielding the higher proportion of treatment participants.

The primary outcome was a rating of functional disability at 180 days post-stroke, measured by the modified Rankin Scale (mRS) (Quinn et al., 2009). The mRS score is ordinal, defined as an integer between 0 (no symptoms) and 6 (death), with lower scores corresponding to improved functioning (Cheng et al., 2014). In the primary analysis comparing treatment to control, the average treatment effect (ATE) was inferred, i.e., the difference in population proportions with 180-day mRS ≤ 3. The estimate of ATE was 0.11 (95% CI: [-0.09, 0.29]), using the 52 treatment and 38 control participants with recorded 180-day mRS scores. Patients and doctors may also be interested in the fraction who benefit, i.e., the fraction of patients who would have a better 180-day mRS under treatment than under control. Since it divides mRS into two categories (≤ 3 or > 3), the ATE misses benefits within a category. In general, the population ATE is not designed to be informative about the fraction. If the outcome is ordinal, the ATE (e.g., the mean difference in the outcome between

treatment and control) can be large, while the fraction is small; this could occur if the majority get zero benefit while a minority have a large benefit.

2.3 Bound Parameters

Denote Y_C and Y_T as the potential outcomes under control and treatment, respectively. Suppose the outcome is ordinal with L levels, i.e., $1, 2, .., L$, ordered from least to most favorable. For MISTIE, we recode mRS score in this way with $L = 7$, setting 1 to represent death, and 7 to represent no symptoms. This definition of mRS score will be used throughout this chapter. Let X be a prognostic baseline variable collected in the randomized trial. For each participant, define the vector including the baseline variable and both potential outcomes as $V = (X, Y_C, Y_T)$. We let P denote a generic joint distribution on (X, Y_C, Y_T) and P_0 denote the true (unknown) distribution on (X, Y_C, Y_T). We assume that each participant's vector V is an independent, identically distributed draw from P_0. For each participant, the observed data is (X, A, Y), where A is the random treatment assignment (1 if treatment, 0 if control) which is independent of V, and Y is the observed outcome corresponding to the treatment assigned, i.e., $Y = AY_T + (1 - A)Y_C$.

Our goal is to learn about the fraction who benefit, i.e., the parameter $\psi = P(Y_T > Y_C)$, despite never observing the full pair of potential outcomes for any participant. Although the fraction ψ is generally non-identifiable, certain possibilities can be ruled

out using the marginal distributions of Y_C and Y_T, which are identifiable. Let F_C, F_T denote the marginal distribution functions of Y_C, Y_T under P; let p_C, p_T denote the corresponding probability mass functions. Section 2.3.1 gives bounds on ψ based on F_C and F_T. The bounds can be improved by incorporating X or assumptions about P, as discussed in Section 2.3.2.

2.3.1 Sharp bounds on the fraction who benefit based on F_C and F_T

Let $\pi_{i,j}$ be the fraction of the population with $(Y_C, Y_T) = (i, j)$, i.e., $\pi_{i,j} = P(Y_C = i, Y_T = j)$. Let $\mathcal{L}$ be the set of integers from 1 to L. The $\pi_{i,j}$'s $(i, j \in \mathcal{L})$ form an L x L matrix giving the joint distribution of the potential outcomes (JDPO), depicted in Figure 2.1 for MISTIE. The row and column sums of the matrix correspond to p_C and p_T, respectively, as shown in the figure. For example, the sum of the $\pi_{i,j}$'s in the fourth column equals $p_T(4)$.

The population can be partitioned into three categories based on potential outcomes (Y_C, Y_T): those for whom assignment to treatment (compared to control) would have no effect $(Y_T = Y_C)$, harm $(Y_T < Y_C)$, or benefit $(Y_T > Y_C)$. These categories correspond to the yellow, red, and green regions in Figure 2.1, respectively. The parameter ψ is the fraction of the population in the green region, i.e., the sum of $\pi_{i,j}$ over indices (i, j) with $j > i$. The value of ψ, in general, is non-identifiable since for

each participant we only observe one component of (Y_C, Y_T), and therefore do not know which of the three regions she/he belongs to.

Let $\psi_l(F_C, F_T)$ and $\psi_u(F_C, F_T)$ denote the sharp lower and upper bounds on ψ, given F_C and F_T, i.e., the best possible bounds for ψ that could be obtained if F_C, F_T were known. We drop the dependence on F_C, F_T for conciseness and simply write ψ_l and ψ_u. We say a joint distribution P' on (Y_C, Y_T) is consistent with the pair of marginal distribution functions F_C, F_T if, under P', the marginal distribution of Y_C equals F_C and the marginal distribution of Y_T equals F_T. The lower bound ψ_l is:

$$\psi_l = \inf\{P'(Y_T > Y_C) : P' \text{ on } (Y_C, Y_T) \text{ consistent with } F_C,\, F_T\} \tag{2.1}$$

$$= \min \left\{ \sum_{\substack{j>i \\ i,j \in \mathcal{L}}} \pi_{i,j} : \begin{array}{l} \pi_{i,j} \geq 0 \text{ for all } i, j \in \mathcal{L} \\ \sum_{i'=1}^{i} \sum_{j=1}^{L} \pi_{i',j} = F_C(i) \text{ for } i = 1, .., L-1 \\ \sum_{j'=1}^{j} \sum_{i=1}^{L} \pi_{i,j'} = F_T(j) \text{ for } j = 1, .., L-1 \\ \sum_{i=1}^{L} \sum_{j=1}^{L} \pi_{i,j} = 1 \end{array} \right\}. \tag{2.2}$$

The upper bound ψ_u is (2.1) with inf replaced by sup, and (2.2) with min replaced by max. In other words, if one were to compute $\sum_{j>i} \pi_{i,j}$ for every possible matrix of $\pi_{i,j}$'s with row sums consistent with p_C and column sums consistent with p_T, ψ_l and ψ_u would be the minimum and maximum. Given the form of (2.2), ψ_l and ψ_u are solutions to linear programs (Borusyak, 2015). The objective function is linear because it is the sum of the $\pi_{i,j}$'s in the green region. The constraints are all linear equality or inequality constraints. In the binary case ($L = 2$), ψ_l and ψ_u simplify to

$\psi_l = \max\{0, p_T(2) - p_C(2)\}$ and $\psi_u = \min\{p_C(1), p_T(2)\}$ (Gadbury et al., 2004). For a continuous outcome, the sharp lower and upper bounds, given only F_C and F_T, have formulae $\sup_{y\in\mathbb{R}}\{F_C(y) - F_T(y)\}$ and $1 - \sup_{y\in\mathbb{R}}\{F_T(y) - F_C(y)\}$, respectively (Williamson and Downs, 1990; Fan and Park, 2010). For ordinal outcomes, the lower bound ψ_l equals $\sup_{y\in\mathbb{R}}\{F_C(y) - F_T(y)\}$, while the upper bound ψ_u can be different from $1 - \sup_{y\in\mathbb{R}}\{F_T(y) - F_C(y)\}$.

Example (Case where $\psi_u \neq 1 - \sup_{y\in\mathbb{R}}\{F_T(y) - F_C(y)\}$) *Consider the case of a binary outcome ($L = 2$) with possible values 1 = failure and 2 = success. Assume the marginal distributions of Y_C and Y_T are identical, with: $p_C(1) = p_C(2) = p_T(1) = p_T(2) = 0.5$. By Gadbury et al. (2004), we have $\psi_u = \min\{p_C(1), p_T(2)\} = 0.5$. In contrast, we have*

$$1 - \sup_{y\in\mathbb{R}}\{F_T(y) - F_C(y)\} \;=\; 1.$$

2.3.2 General formulation of sharp bounds on the fraction who benefit

We generalize the bound formulation to incorporate a baseline variable and support restrictions. Since they offer new information, these features can narrow the bounds (Fan and Park, 2010; Kim, 2014). We consider a baseline, i.e., pre-randomization,

variable that is categorical and conjectured to be prognostic for (i.e., correlated with) the outcome. Suppose the baseline variable X has $K < \infty$ possible values: $x_1, .., x_K$. Let p_X be its probability mass function, with $p_X(x_k) = P(X = x_k) > 0$ for each k. The population can be stratified into K subpopulations, based on X. For each k, let F_C^k and F_T^k be the cumulative distribution functions of Y_C and Y_T conditional on $X = x_k$.

In addition to a baseline variable, support restrictions are another feature that can be incorporated. Support restrictions are assumptions that $P(Y_C = i, Y_T = j) = 0$ for specific (i, j) pairs. They are encoded by a function $g : \mathcal{L} \times \mathcal{L} \to \{0, 1\}$ that maps a potential outcome pair $(Y_C, Y_T) = (i, j)$ to 0 if the pair is assumed not possible, and 1 otherwise. Equivalently, $g(i, j) = 0$ encodes the assumption that $\pi_{i,j} = 0$. The support restrictions in our application (Section 2.5) are restrictions on harm/benefit. The restriction Harm $\leq d$ levels is: $\pi_{i,j} = 0$ if $i - j > d$. The no harm assumption is the special case with $d = 0$. The restriction Benefit $\leq d$ levels is: $\pi_{i,j} = 0$ if $j - i > d$. Figure 2.2 illustrates Benefit ≤ 3 levels for MISTIE. We refer to support restrictions simply as restrictions.

We assume the restrictions, i.e., the function g, are prespecified and known. Let $\mathcal{R}$ be the subclass of joint distributions P' on (Y_C, Y_T) that satisfy the restrictions, i.e., $P'(Y_C = i, Y_T = j) = 0$ if $g(i, j) = 0$.

Assumption 1 *The true joint distribution $P_0(X, Y_C, Y_T)$ is consistent with $\mathcal{R}$, i.e., the distribution $P_0(Y_C, Y_T)$, which is formed by marginalizing $P_0(X, Y_C, Y_T)$ over X,*

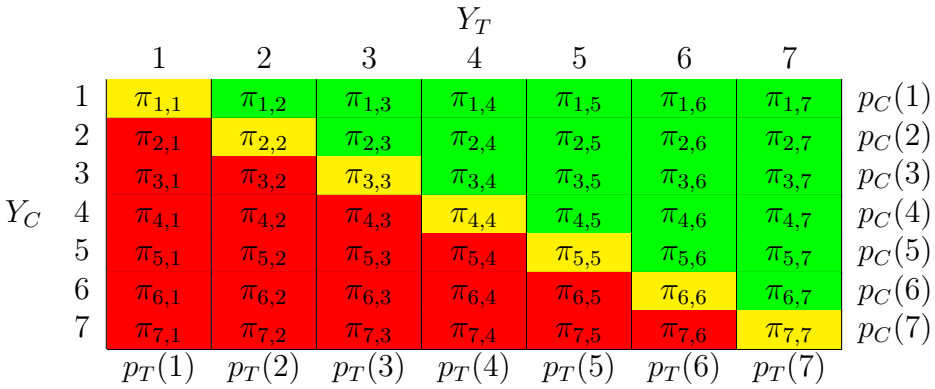


Figure 2.1: Joint Distribution of the Potential Outcomes.

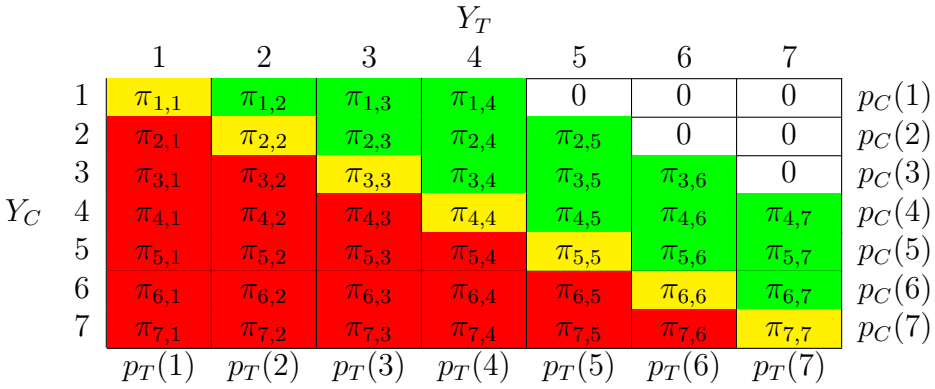


Figure 2.2: Support Restriction: Benefit ≤ 3 levels.

is in $\mathcal{R}$.

Let $\psi_l^{\mathcal{R},X}$ and $\psi_u^{\mathcal{R},X}$ denote the sharp lower and upper bounds on ψ, respectively, given the baseline variable X and restrictions $\mathcal{R}$. These bounds are functions of $\mathcal{R}$ and the identifiable components of $P(X, Y_C, Y_T)$ in a randomized trial where study arm is assigned independent of X, i.e., the components $\{F_C^k, F_T^k\}_{k=1}^K$ and p_X. Formally, we have

$$\psi_l^{\mathcal{R},X} = \psi_l^{\mathcal{R},X}(\{F_C^k, F_T^k\}_{k=1}^K, p_X) \tag{2.3}$$

$$= \inf\left\{P'(Y_T > Y_C) : P' \text{ on } (X, Y_C, Y_T) \text{ consistent with } \mathcal{R}, \{F_C^k, F_T^k\}_{k=1}^K, p_X\right\}. \tag{2.4}$$

The upper bound $\psi_u^{\mathcal{R},X} = \psi_u^{\mathcal{R},X}(\{F_C^k, F_T^k\}_{k=1}^K, p_X)$ is (2.4), with sup in place of inf.

Let $\psi_l^{\mathcal{R}}$ denote the lower bound with restrictions $\mathcal{R}$ but ignoring the baseline variable X, i.e., (2.3)-(2.4) with $\{F_C^k, F_T^k\}_{k=1}^K, p_X$ replaced by F_C, F_T. Let ψ_l^X denote the lower bound with the baseline variable but no restrictions $\mathcal{R}$, i.e., (2.3)-(2.4) with $\mathcal{R}$ omitted. Analogous definitions apply for the upper bounds. The bounds ψ_l, ψ_u from Section 2.3.1 are equivalent to (2.3)-(2.4) with $\mathcal{R}$ omitted and $\{F_C^k, F_T^k\}_{k=1}^K, p_X$ replaced by F_C, F_T. Each of the bounds $\psi_l^{\mathcal{R},X}, \psi_l^{\mathcal{R}}, \psi_l^X, \psi_l$ is a function of the joint distribution $P = P(X, Y_C, Y_T)$ through the corresponding identifiable components. E.g., $\psi_l^{\mathcal{R},X} = \psi_l^{\mathcal{R},X}(P)$ depends on P through the components $\{F_C^k, F_T^k\}_{k=1}^K, p_X$ of P; the bound ψ_l is a function of P through the (less informative) components F_C, F_T.

We suppress the dependence of these parameters on P for conciseness.

Incorporating a baseline variable or restriction leads to a larger or equal lower bound, and smaller or equal upper bound.

Theorem 1 *Consider any restrictions $\mathcal{R}$, baseline variable X, and joint distribution P on (X, Y_C, Y_T) consistent with $\mathcal{R}$. Then (i) $\psi_l^{\mathcal{R},X} \geq \max\{\psi_l^X, \psi_l^{\mathcal{R}}\}$ and $\min\{\psi_l^X, \psi_l^{\mathcal{R}}\} \geq \psi_l$, (ii) $\psi_u^{\mathcal{R},X} \leq \min\{\psi_u^X, \psi_u^{\mathcal{R}}\}$ and $\max\{\psi_u^X, \psi_u^{\mathcal{R}}\} \leq \psi_u$, where each bound parameter is evaluated at P.*

Proof *Let F_C, F_T, and p_X be the marginal distributions of Y_C, Y_T, and X, respectively, under P. For each $k = 1, .., K$, let F_C^k and F_T^k be the conditional distributions of Y_C and Y_T given $X = x_k$, under P. We have the following:*

$$\psi_l^{\mathcal{R},X} = \inf\left\{P'(Y_T > Y_C) : P' \text{ on } (X, Y_C, Y_T) \text{ consistent with } \mathcal{R}, \{F_C^k, F_T^k\}_{k=1}^K, p_X\right\} \tag{2.5}$$

$$\geq \inf\left\{P'(Y_T > Y_C) : P' \text{ on } (X, Y_C, Y_T) \text{ consistent with } \{F_C^k, F_T^k\}_{k=1}^K, p_X\right\} \tag{2.6}$$

$$= \psi_l^X.$$

$$\psi_l^X = \inf\left\{P'(Y_T > Y_C) : P' \text{ on } (X, Y_C, Y_T) \text{ consistent with } \{F_C^k, F_T^k\}_{k=1}^K, p_X\right\} \quad (2.7)$$

$$\geq \inf\left\{P'(Y_T > Y_C) : P' \text{ on } (X, Y_C, Y_T) \text{ consistent with } F_C, F_T\right\} \quad (2.8)$$

$$= \psi_l.$$

$$\psi_l^{\mathcal{R},X} = \inf\left\{P'(Y_T > Y_C) : P' \text{ on } (X, Y_C, Y_T) \text{ consistent with } \mathcal{R}, \{F_C^k, F_T^k\}_{k=1}^K, p_X\right\} \quad (2.9)$$

$$\geq \inf\left\{P'(Y_T > Y_C) : P' \text{ on } (X, Y_C, Y_T) \text{ consistent with } \mathcal{R}, F_C, F_T\right\} \quad (2.10)$$

$$= \psi_l^{\mathcal{R}}.$$

$$\psi_l^{\mathcal{R}} = \inf\left\{P'(Y_T > Y_C) : P' \text{ on } (X, Y_C, Y_T) \text{ consistent with } \mathcal{R}, F_C, F_T\right\} \quad (2.11)$$

$$\geq \inf\left\{P'(Y_T > Y_C) : P' \text{ on } (X, Y_C, Y_T) \text{ consistent with } F_C, F_T\right\} \quad (2.12)$$

$$= \psi_l.$$

All of these bound parameters are well-defined since P is consistent with F_C, F_T, $\{F_C^k, F_T^k\}_{k=1}^K$, p_X, and $\mathcal{R}$. Inequalities (2.6) *and* (2.10) *hold because the constraint that P' is consistent with $\mathcal{R}$ has been removed on the right sides, so the* inf *is being taken over an increased set leading to a smaller or equal value. Inequalities* (2.8) *and* (2.12) *hold because the* inf *is being taken over an increased set on the right sides, since any P' consistent with all $\{F_C^k, F_T^k\}_{k=1}^K, p_X$ is also consistent with F_C, F_T, as shown*

below:

If P' consistent with all $\{F_C^k, F_T^k\}_{k=1}^K, p_X$ (which are the marginal distributions corresponding to P), then for any y,

$$\begin{aligned} P'(Y_C \leq y) &= \sum_{k=1}^{K} P'(Y_C \leq y | X = x_k) P'(X = x_k) \\ &= \sum_{k=1}^{K} F_C^k(y) p_X(x_k) \\ &= \sum_{k=1}^{K} P(Y_C \leq y | X = x_k) P(X = x_k) \\ &= P(Y_C \leq y) \\ &= F_C(y), \end{aligned}$$

so P' is consistent with F_C. Using an analogous proof, it can be shown that P' is consistent with F_T.

Since $\psi_l^{\mathcal{R},X} \geq \psi_l^X$ and $\psi_l^{\mathcal{R},X} \geq \psi_l^{\mathcal{R}}$, we have $\psi_l^{\mathcal{R},X} \geq \max\{\psi_l^X, \psi_l^{\mathcal{R}}\}$. Since $\psi_l^X \geq \psi_l$ and $\psi_l^{\mathcal{R}} \geq \psi_l$, we have $\min\{\psi_l^X, \psi_l^{\mathcal{R}}\} \geq \psi_l$. The results for the upper bound parameters can be proved analogously. □

Just as in Fan and Park (2010), the baseline variable X will not affect the bounds if it is independent of (Y_C, Y_T). This claim is justified at the end of this section.

Restrictions $\mathcal{R}$ may be inconsistent with a set of marginal distributions $\{F_C^k, F_T^k\}_{k=1}^K$, p_X, i.e., there may not exist a joint distribution P' on (X, Y_C, Y_T) that is consistent with $\mathcal{R}$ and $\{F_C^k, F_T^k\}_{k=1}^K, p_X$. In this case, the bound parameter (2.4) is undefined,

since the set of distributions on the right hand side is empty. This cannot occur if the distribution P is consistent with $\mathcal{R}$. However, the user may impose restrictions $\mathcal{R}$ that are in violation of Assumption 1. This can lead to the bound parameters evaluated at P_0, such as $\psi_l^{\mathcal{R},X}(P_0)$, being undefined.

Bounds on the fraction who benefit can be derived for a subpopulation. For any given k, let $\psi_{l,k}^{\mathcal{R},X}$ and $\psi_{u,k}^{\mathcal{R},X}$ denote the sharp lower and upper bounds for subpopulation k, given $\{F_C^k, F_T^k\}$ and the restrictions $\mathcal{R}$. The lower bound $\psi_{l,k}^{\mathcal{R},X}$ is:

$$\psi_{l,k}^{\mathcal{R},X} = \inf\{P'(Y_T > Y_C | X = x_k) : P' \text{ on } (X, Y_C, Y_T) \text{ consistent with } \mathcal{R}, F_C^k, F_T^k\}.$$

(2.13)

$$= \min \left\{ \sum_{\substack{j>i \\ i,j \in \mathcal{L}}} \pi_{i,j}^k : \begin{array}{l} \pi_{i,j}^k \geq 0 \text{ for all } i,j \in \mathcal{L} \\ \sum_{i'=1}^{i} \sum_{j=1}^{L} \pi_{i',j}^k = F_C^k(i) \text{ for all } i = 1,..,L-1 \\ \sum_{j'=1}^{j} \sum_{i=1}^{L} \pi_{i,j'}^k = F_T^k(j) \text{ for all } j = 1,..,L-1 \\ \sum_{i=1}^{L} \sum_{j=1}^{L} \pi_{i,j}^k = 1 \\ \pi_{i,j}^k = 0 \text{ if } g(i,j) = 0 \end{array} \right\}. \quad (2.14)$$

In (2.14), we let $\pi_{i,j}^k = P'(Y_C = i, Y_T = j | X = x_k)$ for each i, j. The upper bound $\psi_{u,k}^{\mathcal{R},X}$ is (2.13) with sup in place of inf, and (2.14) with max in place of min. The population lower bound is a weighted sum of the subpopulation lower bounds, where the weights

equal the size of the subpopulation relative to the full population. Analogously, the population upper bound is a weighted sum of the subpopulation upper bounds, with the same weights.

Theorem 2 *Consider any restriction $\mathcal{R}$, baseline variable X, and joint distribution P on (X, Y_C, Y_T) that is consistent with $\mathcal{R}$. Then we have*

$$\psi_l^{\mathcal{R},X} = \sum_{k=1}^{K} \psi_{l,k}^{\mathcal{R},X} p_X(x_k), \qquad \psi_u^{\mathcal{R},X} = \sum_{k=1}^{K} \psi_{u,k}^{\mathcal{R},X} p_X(x_k),$$

where each parameter is evaluated at P. This also holds with the restrictions $\mathcal{R}$ omitted.

Proof *The bounds $\psi_l^{\mathcal{R},X}$ and $\psi_u^{\mathcal{R},X}$ are defined as:*

$$\begin{aligned}
\psi_l^{\mathcal{R},X} &= \inf\left\{P'(Y_T > Y_C) : P' \text{ on } (X, Y_C, Y_T) \text{ consistent with } \mathcal{R}, \{F_C^k, F_T^k\}_{k=1}^K, p_X\right\}, \\
\psi_u^{\mathcal{R},X} &= \sup\left\{P'(Y_T > Y_C) : P' \text{ on } (X, Y_C, Y_T) \text{ consistent with } \mathcal{R}, \{F_C^k, F_T^k\}_{k=1}^K, p_X\right\},
\end{aligned}$$

where $F_C^k(y) = P(Y_C \leq y | X = x_k)$ and $F_T^k(y) = P(Y_T \leq y | X = x_k)$ for any y and k, and $p_X(x_k) = P(X = x_k)$ for any k. For each k, we have that:

$$\begin{aligned}
\psi_{l,k}^{\mathcal{R},X} &= \inf\left\{P'(Y_T > Y_C | X = x_k) : P' \text{ on } (X, Y_C, Y_T) \text{ consistent with } \mathcal{R}, F_C^k, F_T^k\right\}, \\
\psi_{u,k}^{\mathcal{R},X} &= \sup\left\{P'(Y_T > Y_C | X = x_k) : P' \text{ on } (X, Y_C, Y_T) \text{ consistent with } \mathcal{R}, F_C^k, F_T^k\right\},
\end{aligned}$$

Since $\psi_l^{\mathcal{R},X}$, $\psi_u^{\mathcal{R},X}$, $\{\psi_{l,k}^{\mathcal{R},X}\}_{k=1}^K$, and $\{\psi_{u,k}^{\mathcal{R},X}\}_{k=1}^K$ are solutions to linear programs, inf

can be replaced with min, *and* sup *with* max *in the four definitions above.*

We now show that $\sum_{k=1}^{K} \psi_{l,k}^{\mathcal{R},X} p_X(x_k)$ *and* $\sum_{k=1}^{K} \psi_{u,k}^{\mathcal{R},X} p_X(x_k)$ *are in the set*

$$\left\{P'(Y_T > Y_C) : P' \text{ on } (X, Y_C, Y_T) \text{ consistent with } \mathcal{R}, \{F_C^k, F_T^k\}_{k=1}^K, p_X\right\}. \quad (2.15)$$

That is, there exist joint distributions P_l *and* P_u *on* (X, Y_C, Y_T) *that (i) are consistent with* $\mathcal{R}$*, (ii) are consistent with* $\{F_C^k, F_T^k\}_{k=1}^K$ *and* p_X*, and (iii)* $P_l(Y_T > Y_C) = \sum_{k=1}^{K} \psi_{l,k}^{\mathcal{R},X} p_X(x_k)$*,* $P_u(Y_T > Y_C) = \sum_{k=1}^{K} \psi_{u,k}^{\mathcal{R},X} p_X(x_k)$*. For any* k*, there exists a joint distribution* $P_{l,k}$ *on* (X, Y_C, Y_T) *that is consistent with* $\mathcal{R}$*,* F_C^k*, and* F_T^k*, and satisfies* $P_{l,k}(Y_T > Y_C | X = x_k) = \psi_{l,k}^{\mathcal{R},X}$*, by the definition of* $\psi_{l,k}^{\mathcal{R},X}$*. Define* P_l *as follows. Let* $P_l(Y_C, Y_T | X = x_k) = P_{l,k}(Y_C, Y_T | X = x_k)$ *for each* k*. Let* $P_l(X = x_k) = p_X(x_k)$ *for each* k*. It follows that* P_l *satisfies (i) and (ii). Also,* $P_l(Y_T > Y_C) = \sum_{k=1}^{K} P_l(Y_T > Y_C | X = x_k) P_l(X = x_k) = \sum_{k=1}^{K} \psi_{l,k}^{\mathcal{R},X} p_X(x_k)$*. The joint distribution* P_u *can be defined analogously as* P_l*. Thus,* $\sum_{k=1}^{K} \psi_{l,k}^{\mathcal{R},X} p_X(x_k)$ *and* $\sum_{k=1}^{K} \psi_{u,k}^{\mathcal{R},X} p_X(x_k)$ *are in* (2.15)*.*

Now it remains to show that $\sum_{k=1}^{K} \psi_{l,k}^{\mathcal{R},X} p_X(x_k)$ *and* $\sum_{k=1}^{K} \psi_{u,k}^{\mathcal{R},X} p_X(x_k)$ *are the minimum and maximum of* (2.15)*, respectively. We do a proof by contradiction. Suppose that the minimum of* (2.15) *is smaller than* $\sum_{k=1}^{K} \psi_{l,k}^{\mathcal{R},X} p_X(x_k)$*. This would imply that, for some* k*, there exists a distribution* P' *on* $(Y_C, Y_T | X = x_k)$ *that is consistent with* $\mathcal{R}, F_C^k, F_T^k$ *and with* $P'(Y_T > Y_C | X = x_k) < \psi_{l,k}^{\mathcal{R},X}$*. However, this contradicts the definition of* $\psi_{l,k}^{\mathcal{R},X}$*. Thus,* $\sum_{k=1}^{K} \psi_{l,k}^{\mathcal{R},X} p_X(x_k)$ *is the minimum of* (2.15)*.*

It can be shown analogously that $\sum_{k=1}^{K} \psi_{u,k}^{\mathcal{R},X} p_X(x_k)$ *is the maximum of* (2.15). *We conclude that*

$$\psi_l^{\mathcal{R},X} = \sum_{k=1}^{K} \psi_{l,k}^{\mathcal{R},X} p_X(x_k), \qquad \psi_u^{\mathcal{R},X} = \sum_{k=1}^{K} \psi_{u,k}^{\mathcal{R},X} p_X(x_k).$$

The relationship also holds if the restrictions $\mathcal{R}$ *are omitted. This can be proved as above, with* $\mathcal{R}$ *corresponding to the indicator function* g *that maps all potential outcome pairs to 1 (which is equivalent to no restrictions).* □

Theorem 2 is a useful result since it shows that, to incorporate a baseline variable, one only needs to compute the bounds within subpopulation and take a weighted sum of the subpopulation bounds. In Section 2.4, we also use this strategy to estimate the bounds that incorporate a baseline variable. In other words, we compute the bound estimate within each stratum and take the weighted sum of the bound estimates. Earlier in this section, it was claimed that a baseline variable X that is independent of (Y_C, Y_T) does not result in any narrowing of the bounds. This claim follows from Theorem 2. If X is independent of (Y_C, Y_T), we have $F_C^k = F_C$ and $F_T^k = F_T$ for all k. This implies that $\psi_{l,k}^{\mathcal{R},X} = \psi_l^{\mathcal{R}}$ and $\psi_{u,k}^{\mathcal{R},X} = \psi_u^{\mathcal{R}}$. By Theorem 2, we have $\psi_l^{\mathcal{R},X} = \psi_l^{\mathcal{R}}$ and $\psi_u^{\mathcal{R},X} = \psi_u^{\mathcal{R}}$.

2.4 Bound Estimators

We discuss estimators for the bound parameters defined in Section 2.3, using data from a randomized trial with n participants. We make the assumption below:

Assumption 2 *(i) For each participant m, her/his vector $V_m = (X_m, Y_{C,m}, Y_{T,m})$ is an independent, identically distributed draw from P_0. (ii) The treatment assignments, A_m, $m = 1, .., n$, are independent, identically distributed Bernoulli(θ), and are independent of $\{V_m\}_{m=1}^n$. (iii) The observed data vector for participant m is (X_m, A_m, Y_m) where $Y_m = A_m Y_{T,m} + (1 - A_m) Y_{C,m}$.*

Above, (ii) is justified by randomization and we assume the randomization probability $\theta : 0 < \theta < 1$. The equality in (iii) is called the consistency assumption, which connects potential outcomes (Y_C, Y_T) and treatment assignment A to the observed outcome Y.

2.4.1 Plug-in estimator

One might consider a plug-in (also called substitution) estimator, where in place of p_X, $\{F_C^k, F_T^k\}_{k=1}^K$, (or of F_C, F_T), the following sample proportions are used:

$$\begin{aligned}
\widehat{p}_X(x_k) &= \frac{1}{n}\sum_{m=1}^{n} 1(X_m = x_k), \\
\widehat{F}_C^k(y) &= \frac{\sum_{m=1}^n 1(Y_m \le y, A_m = 0, X_m = x_k)}{\sum_{m=1}^n 1(A_m = 0, X_m = x_k)}, \\
\widehat{F}_T^k(y) &= \frac{\sum_{m=1}^n 1(Y_m \le y, A_m = 1, X_m = x_k)}{\sum_{m=1}^n 1(A_m = 1, X_m = x_k)},
\end{aligned}$$

for any y and k. Define $\widehat{F}_C, \widehat{F}_T$ as in the above display, with $X_m = x_k$ omitted. Above, $1(E)$ has value 1 if E occurs and 0 otherwise. We use the hat symbol to denote plug-in estimators, e.g., $\widehat{\psi}_l$.

The plug-in estimator can be inconsistent when support restrictions are made, even if they are correct. Consider the case where the outcome is binary, the baseline variable is ignored, and the true, unknown joint distribution P_0 on (Y_C, Y_T) satisfies $P_0(Y_C = Y_T) = 1$. Then the true marginals satisfy $P_0(Y_C = y) = P_0(Y_T = y)$ for each $y \in \{1, 2\}$. Let the restrictions $\mathcal{R}$ represent no harm, i.e., the event $(Y_C = 2, Y_T = 1)$ is assumed to have probability 0. The restrictions are consistent with P_0. The bound parameters at P_0 satisfy $(\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}) = (0, 0)$. Let the randomization probability θ be 0.5. If $\widehat{F}_T(1) > \widehat{F}_C(1)$, no joint distribution on (Y_C, Y_T) is consistent with both $\widehat{F}_C, \widehat{F}_T$ and $\mathcal{R}$; in this case, $\widehat{\psi}_l^{\mathcal{R}}$ and $\widehat{\psi}_u^{\mathcal{R}}$ are undefined. The probability $P_0(\widehat{F}_T(1) > \widehat{F}_C(1))$ converges to 0.5 as n goes to infinity, as proved below.

Proof *Consider any n in $\mathbb{N}$. For any given randomized trial, we have either $\widehat{F}_T(1) > \widehat{F}_C(1)$ or $\widehat{F}_T(1) < \widehat{F}_C(1)$ or $\widehat{F}_T(1) = \widehat{F}_C(1)$. Since $\theta = \frac{1}{2}$ and $F_C(1) = F_T(1)$, we have $P_0(\widehat{F}_T(1) > \widehat{F}_C(1)) = P_0(\widehat{F}_T(1) < \widehat{F}_C(1))$. Therefore,*

$$P_0(\widehat{F}_T(1) > \widehat{F}_C(1)) = \frac{1 - P_0(\widehat{F}_T(1) = \widehat{F}_C(1))}{2}.$$

Now let n go to infinity. Then we have

$$\lim_{n\to\infty} P_0(\widehat{F}_T(1) > \widehat{F}_C(1)) = \lim_{n\to\infty} \frac{1 - P_0(\widehat{F}_T(1) = \widehat{F}_C(1))}{2} = \frac{1}{2}.$$

□

Because $P_0(\widehat{F}_T(1) > \widehat{F}_C(1))$ converges to 0.5 as n goes to infinity, $\widehat{\psi}_l^{\mathcal{R}}$ and $\widehat{\psi}_u^{\mathcal{R}}$ are undefined with approximately 0.5 probability for arbitrarily large n.

In general, $\widehat{\psi}_l^{\mathcal{R}}$ and $\widehat{\psi}_u^{\mathcal{R}}$ are inconsistent if the linear programs for $\psi_l^{\mathcal{R}}$ and $\psi_u^{\mathcal{R}}$ are feasible but an arbitrarily small perturbation to F_C and F_T could make them infeasible. (The linear programs for $\psi_l^{\mathcal{R}}$ and $\psi_u^{\mathcal{R}}$ are like (2.2), except with the constraint "$\pi_{i,j} = 0$ if $g(i,j) = 0$" included.) Analogously, $\widehat{\psi}_l^{\mathcal{R},X}$ and $\widehat{\psi}_u^{\mathcal{R},X}$ are inconsistent if, for some k, the linear programs for $\psi_{l,k}^{\mathcal{R},X}$ and $\psi_{u,k}^{\mathcal{R},X}$ given by (2.14) are feasible but an arbitrarily small change to F_C^k and F_T^k can make them infeasible. We refer to these cases as boundary cases. Boundary cases can only occur if restrictions are made. They can occur when the true fraction who benefit and the bound parameters are nonzero, as shown in the following example.

Example *Consider an ordinal outcome with three levels (1,2,3, ordered from worst to best). Take the case where the true joint distribution* P_0 *on* (Y_C, Y_T) *is:*

		Y_T = 1	2	3	
	1	0.2	0.3	0	0.5
Y_C	2	0	0.3	0	0.3
	3	0	0	0.2	0.2
		0.2	0.6	0.2	

Therefore, the fraction who benefit $\psi = P_0(Y_T > Y_C)$ *is 0.3.*

Suppose that the no harm assumption is imposed, and thus $\mathcal{R} = \{no\ harm\}$*. The only joint distribution on* (Y_C, Y_T) *that satisfies the no harm assumption and has the same marginal distributions as* P_0 *is* P_0 *itself. Hence,* $\psi_l^{\mathcal{R}} = \psi_u^{\mathcal{R}} = 0.3$*.*

Consider the plug-in estimators, $\widehat{\psi}_l^{\mathcal{R}}$ *and* $\widehat{\psi}_u^{\mathcal{R}}$*. These are the minimum and maximum fraction who benefit, among all matrices that satisfy the no harm assumption and have marginal cumulative distribution functions* $\widehat{F}_C$ *and* $\widehat{F}_T$*. The structure of such matrices is depicted in the figure directly below. (Having marginal cumulative distribution functions* $\widehat{F}_C$ *and* $\widehat{F}_T$ *is equivalent to having marginal probability mass functions* $\widehat{p}_C$ *and* $\widehat{p}_T$*.)*

		Y_T = 1	2	3	
Y_C	1	$\pi_{1,1}$	$\pi_{1,2}$	$\pi_{1,3}$	$\widehat{p}_C(1)$
	2	0	$\pi_{2,2}$	$\pi_{2,3}$	$\widehat{p}_C(2)$
	3	0	0	$\pi_{3,3}$	$\widehat{p}_C(3)$
		$\widehat{p}_T(1)$	$\widehat{p}_T(2)$	$\widehat{p}_T(3)$	

There are two conditions under which $\widehat{\psi}_l^{\mathcal{R}}$ and $\widehat{\psi}_u^{\mathcal{R}}$ are undefined (i.e., there exist no matrices with the above form). These conditions are $\widehat{p}_C(3) > \widehat{p}_T(3)$ and $\widehat{p}_T(1) > \widehat{p}_C(1)$.

Suppose $\theta = 1/2$. As n goes to ∞, the probability of the event $\widehat{p}_T(1) > \widehat{p}_C(1)$ converges to 0 because $p_T(1) = 0.2$ while $p_C(1) = 0.5$. However, as n goes to ∞, the probability of $\widehat{p}_C(3) > \widehat{p}_T(3)$ converges to 1/2 because $p_C(3) = p_T(3) = 0.2$. Thus, at arbitrarily large n, the plug-in estimators $\widehat{\psi}_l^{\mathcal{R}}$ and $\widehat{\psi}_u^{\mathcal{R}}$ are undefined with probability approximately 1/2. Hence, $\widehat{\psi}_l^{\mathcal{R}}$ and $\widehat{\psi}_u^{\mathcal{R}}$ do not converge in probability to $\psi_l^{\mathcal{R}}$ and $\psi_u^{\mathcal{R}}$.

2.4.2 Proposed estimator

Our estimators of the parameters $\psi_l^{\mathcal{R},X}$ and $\psi_u^{\mathcal{R},X}$, respectively, are defined as

$$\overline{\psi}_l^{\mathcal{R},X} = \sum_{k=1}^{K} \overline{\psi}_{l,k}^{\mathcal{R},X} \widehat{p}_X(x_k), \qquad \overline{\psi}_u^{\mathcal{R},X} = \sum_{k=1}^{K} \overline{\psi}_{u,k}^{\mathcal{R},X} \widehat{p}_X(x_k), \tag{2.16}$$

For each k, the term $\overline{\psi}_{l,k}^{\mathcal{R},X}$ is computed by the following sequence of two linear

programs:

$$\overline{\epsilon}_k = \min \left\{ \epsilon_k \geq 0 : \begin{array}{l} \pi^k_{i,j} \geq 0 \text{ for all } i,j \in \mathcal{L} \\ |\sum_{i'=1}^{i} \sum_{j=1}^{L} \pi^k_{i',j} - \widehat{F}^k_C(i)| \leq \epsilon_k \text{ for } i = 1,..,L-1 \\ |\sum_{j'=1}^{j} \sum_{i=1}^{L} \pi^k_{i,j'} - \widehat{F}^k_T(j)| \leq \epsilon_k \text{ for } j = 1,..,L-1 \\ \sum_{i=1}^{L} \sum_{j=1}^{L} \pi^k_{i,j} = 1 \\ \pi^k_{i,j} = 0 \text{ if } g(i,j) = 0 \end{array} \right\} \tag{2.17}$$

$$\overline{\psi}^{\mathcal{R},X}_{l,k} = \min \left\{ \sum_{\substack{j>i \\ i,j \in \mathcal{L}}} \pi^k_{i,j} : \begin{array}{l} \pi^k_{i,j} \geq 0 \text{ for all } i,j \in \mathcal{L} \\ |\sum_{i'=1}^{i} \sum_{j=1}^{L} \pi^k_{i',j} - \widehat{F}^k_C(i)| \leq \overline{\epsilon}_k \text{ for } i = 1,..,L-1 \\ |\sum_{j'=1}^{j} \sum_{i=1}^{L} \pi^k_{i,j'} - \widehat{F}^k_T(j)| \leq \overline{\epsilon}_k \text{ for } j = 1,..,L-1 \\ \sum_{i=1}^{L} \sum_{j=1}^{L} \pi^k_{i,j} = 1 \\ \pi^k_{i,j} = 0 \text{ if } g(i,j) = 0 \end{array} \right\}. \tag{2.18}$$

The term $\overline{\psi}^{\mathcal{R},X}_{u,k}$ is (2.18), with min replaced by max. (2.17) and (2.18) are linear programs because each absolute value statement can be converted to a pair of linear inequalities.

The key idea in (2.18) is that we relaxed the constraint that the marginal distribution functions corresponding to $\pi^k_{i,j}$ equal the empirical marginal distribution

functions in stratum k; we instead allow these to differ by at most $\overline{\epsilon}_k$. As defined in (2.17), the value of $\overline{\epsilon}_k$ is the minimum value that allows the linear programs for $\overline{\psi}_{l,k}^{\mathcal{R},X}$ and $\overline{\psi}_{u,k}^{\mathcal{R},X}$ to be feasible. If the plug-in estimators $\widehat{\psi}_{l,k}^{\mathcal{R},X}$ and $\widehat{\psi}_{u,k}^{\mathcal{R},X}$ are well-defined, we have $\overline{\epsilon}_k = 0$ and thus $\overline{\psi}_{l,k}^{\mathcal{R},X} = \widehat{\psi}_{l,k}^{\mathcal{R},X}$ and $\overline{\psi}_{u,k}^{\mathcal{R},X} = \widehat{\psi}_{u,k}^{\mathcal{R},X}$. If $\widehat{\psi}_{l,k}^{\mathcal{R},X}$ and $\widehat{\psi}_{u,k}^{\mathcal{R},X}$ are undefined, we have $\overline{\epsilon}_k > 0$ allowing $\overline{\psi}_{l,k}^{\mathcal{R},X}$ and $\overline{\psi}_{u,k}^{\mathcal{R},X}$ to be well-defined.

Our estimators that ignore baseline variables and/or have no restrictions, e.g., $\overline{\psi}_l^{\mathcal{R}}$, are defined analogously. Their definitions are shown below. In the case where no restrictions are made, our estimator is equivalent to the plug-in estimator.

With neither baseline variable nor restriction, $\overline{\psi}_l$ and $\overline{\psi}_u$

The lower bound estimator $\overline{\psi}_l$ is computed using a sequence of two linear programs.

$$\overline{\epsilon} = \min \left\{ \epsilon \geq 0 : \begin{array}{l} \pi_{i,j} \geq 0 \text{ for all } i,j \in \mathcal{L} \\ |\sum_{i'=1}^{i} \sum_{j=1}^{L} \pi_{i',j} - \widehat{F}_C(i)| \leq \epsilon \text{ for } i = 1,..,L-1 \\ |\sum_{j'=1}^{j} \sum_{i=1}^{L} \pi_{i,j'} - \widehat{F}_T(j)| \leq \epsilon \text{ for } j = 1,..,L-1 \\ \sum_{i=1}^{L} \sum_{j=1}^{L} \pi_{i,j} = 1 \end{array} \right\}, \quad (2.19)$$

$$\overline{\psi}_l = \min \left\{ \sum_{\substack{j>i \\ i,j \in \mathcal{L}}} \pi_{i,j} : \begin{array}{l} \pi_{i,j} \geq 0 \text{ for all } i,j \in \mathcal{L} \\ |\sum_{i'=1}^{i} \sum_{j=1}^{L} \pi_{i',j} - \widehat{F}_C(i)| \leq \bar{\epsilon} \text{ for } i = 1,..,L-1 \\ |\sum_{j'=1}^{j} \sum_{i=1}^{L} \pi_{i,j'} - \widehat{F}_T(j)| \leq \bar{\epsilon} \text{ for } j = 1,..,L-1 \\ \sum_{i=1}^{L} \sum_{j=1}^{L} \pi_{i,j} = 1 \end{array} \right\}. \quad (2.20)$$

The upper bound estimator $\overline{\psi}_u$ is (2.20), with min replaced by max.

With restriction and ignoring baseline variable, $\overline{\psi}_l^{\mathcal{R}}$ and $\overline{\psi}_u^{\mathcal{R}}$

The lower bound estimator $\overline{\psi}_l^{\mathcal{R}}$ is computed using a sequence of two linear programs:

$$\bar{\epsilon} = \min \left\{ \epsilon \geq 0 : \begin{array}{l} \pi_{i,j} \geq 0 \text{ for all } i,j \in \mathcal{L} \\ |\sum_{i'=1}^{i} \sum_{j=1}^{L} \pi_{i',j} - \widehat{F}_C(i)| \leq \epsilon \text{ for } i = 1,..,L-1 \\ |\sum_{j'=1}^{j} \sum_{i=1}^{L} \pi_{i,j'} - \widehat{F}_T(j)| \leq \epsilon \text{ for } j = 1,..,L-1 \\ \sum_{i=1}^{L} \sum_{j=1}^{L} \pi_{i,j} = 1 \\ \pi_{i,j} = 0 \text{ if } g(i,j) = 0 \end{array} \right\}, \quad (2.21)$$

$$\overline{\psi}_l^{\mathcal{R}} = \min \left\{ \sum_{\substack{j>i \\ i,j \in \mathcal{L}}} \pi_{i,j} : \begin{array}{l} \pi_{i,j} \geq 0 \text{ for all } i,j \in \mathcal{L} \\ |\sum_{i'=1}^{i} \sum_{j=1}^{L} \pi_{i',j} - \widehat{F}_C(i)| \leq \bar{\epsilon} \text{ for } i = 1,..,L-1 \\ |\sum_{j'=1}^{j} \sum_{i=1}^{L} \pi_{i,j'} - \widehat{F}_T(j)| \leq \bar{\epsilon} \text{ for } j = 1,..,L-1 \\ \sum_{i=1}^{L} \sum_{j=1}^{L} \pi_{i,j} = 1 \\ \pi_{i,j} = 0 \text{ if } g(i,j) = 0 \end{array} \right\}. \tag{2.22}$$

The upper bound estimator $\overline{\psi}_u^{\mathcal{R}}$ is (2.22), with min replaced by max.

With baseline variable and no restriction, $\overline{\psi}_l^X$ and $\overline{\psi}_u^X$

The lower and upper bound estimators $\overline{\psi}_l^X$ and $\overline{\psi}_u^X$ are defined as follows:

$$\overline{\psi}_l^X = \sum_{k=1}^{K} \overline{\psi}_{l,k}^X \widehat{p}_X(x_k), \qquad \overline{\psi}_u^X = \sum_{k=1}^{K} \overline{\psi}_{u,k}^X \widehat{p}_X(x_k), \tag{2.23}$$

where the term $\overline{\psi}_{l,k}^X$ is computed using a sequence of two linear programs:

$$\bar{\epsilon}_k = \min \left\{ \epsilon_k \geq 0 : \begin{array}{l} \pi_{i,j}^k \geq 0 \text{ for all } i,j \in \mathcal{L} \\ |\sum_{i'=1}^{i} \sum_{j=1}^{L} \pi_{i',j}^k - \widehat{F}_C^k(i)| \leq \epsilon_k \text{ for } i = 1,..,L-1 \\ |\sum_{j'=1}^{j} \sum_{i=1}^{L} \pi_{i,j'}^k - \widehat{F}_T^k(j)| \leq \epsilon_k \text{ for } j = 1,..,L-1 \\ \sum_{i=1}^{L} \sum_{j=1}^{L} \pi_{i,j}^k = 1 \end{array} \right\}, \tag{2.24}$$

$$\overline{\psi}^X_{l,k} = \quad \min \left\{ \sum_{\substack{j>i \\ i,j\in\mathcal{L}}} \pi^k_{i,j} : \begin{array}{l} \pi^k_{i,j} \geq 0 \text{ for all } i,j \in \mathcal{L} \\ |\sum_{i'=1}^{i}\sum_{j=1}^{L}\pi^k_{i',j} - \widehat{F}^k_C(i)| \leq \overline{\epsilon}_k \text{ for } i = 1,..,L-1 \\ |\sum_{j'=1}^{j}\sum_{i=1}^{L}\pi^k_{i,j'} - \widehat{F}^k_T(j)| \leq \overline{\epsilon}_k \text{ for } j = 1,..,L-1 \\ \sum_{i=1}^{L}\sum_{j=1}^{L}\pi^k_{i,j} = 1 \end{array} \right\}. \tag{2.25}$$

The term $\overline{\psi}^X_{u,k}$ is (2.25), with min replaced by max.

2.4.3 Consistency of the Proposed Estimator

The proposed estimators $\overline{\psi}_l^{\mathcal{R},X}$ and $\overline{\psi}_u^{\mathcal{R},X}$ are consistent, i.e., they converge to the corresponding bound parameters as n goes to infinity. By a similar proof, the estimators that ignore baseline variables and/or have no restrictions are consistent.

Theorem 3 *For any X and $\mathcal{R}$, if P_0 is consistent with $\mathcal{R}$, then $\overline{\psi}_l^{\mathcal{R},X}$ and $\overline{\psi}_u^{\mathcal{R},X}$ are consistent estimators of $\psi_l^{\mathcal{R},X}(P_0)$ and $\psi_u^{\mathcal{R},X}(P_0)$, respectively.*

Proof *In this proof, limits are taken as $n \to \infty$. For example, $\xrightarrow{p}$ signifies convergence in probability as $n \to \infty$. Define 1_E to be a random variable which takes the value 1 if the event E occurs, and 0 otherwise. Generally, define $[z_{i,j}]_{i=1}^L = [z_{1,j}, z_{2,j}, ..., z_{L,j}]$ and $[z_{i,j}]_{j=1}^L = [z_{i,1}, z_{i,2}, ..., z_{i,L}]$. Also, define $\|y\|_2 = \sqrt{\sum_{i=1}^M y_i^2}$, where $y = (y_1, y_2, .., y_M)^T$.*

Consider any X and $\mathcal{R}$. Let P_0 be any joint distribution on (X, Y_C, Y_T) consistent

with $\mathcal{R}$. Suppose that Assumption 2 holds. We will show that $\overline{\psi}_l^{\mathcal{R},X} \xrightarrow{p} \psi_l^{\mathcal{R},X}(P_0)$ and $\overline{\psi}_u^{\mathcal{R},X} \xrightarrow{p} \psi_u^{\mathcal{R},X}(P_0)$ as $n \to \infty$. Below, we suppress the dependence of the bound parameters on P_0 for conciseness. Let $\{F_C^k, F_T^k\}_{k=1}^K, p_X$ be the marginal distributions corresponding to P_0.

The structure of the proof is as follows. We show that: (I) $\hat{p}_X(x_k) \xrightarrow{p} p_X(x_k)$ for each k; and (II) $\overline{\psi}_{l,k}^{\mathcal{R},X} \xrightarrow{p} \psi_{l,k}^{\mathcal{R},X}$ and $\overline{\psi}_{u,k}^{\mathcal{R},X} \xrightarrow{p} \psi_{u,k}^{\mathcal{R},X}$ for each k. By Slutsky's Theorem, (I) and (II) imply that $\overline{\psi}_l^{\mathcal{R},X} \xrightarrow{p} \psi_l^{\mathcal{R},X}$ and $\overline{\psi}_u^{\mathcal{R},X} \xrightarrow{p} \psi_u^{\mathcal{R},X}$. By the Weak Law of Large Numbers, (I) follows from Assumption 2(i). We must show that (II) also holds.

Choose any $k = 1, .., K$. We now prove that $\overline{\psi}_{l,k}^{\mathcal{R},X} \xrightarrow{p} \psi_{l,k}^{\mathcal{R},X}$. Fix $n \in \mathbb{N}$. Consider the following linear program, which we refer to as LP:

$$\min_x c^T x, \text{ subject to } Ax \leq b, x \geq 0.$$

Define x and c as:

$$\begin{aligned} x &= \left[[\pi_{i,1}^k]_{i=1}^L, \ [\pi_{i,2}^k]_{i=1}^L, \ \ldots, \ [\pi_{i,L}^k]_{i=1}^L \right]^T, \\ c^T &= \left[[1_{1>i}]_{i=1}^L, \ [1_{2>i}]_{i=1}^L, \ \ldots, \ [1_{L>i}]_{i=1}^L \right], \end{aligned}$$

where $1_{j>i}$ takes the value 1 if $j > i$ and 0 otherwise. Define A as the matrix that

satisfies:

$$Ax = \Bigg[\sum_{g(i,j)=0} \pi^k_{i,j}, \quad \sum_{i=1}^{L}\sum_{j=1}^{L} \pi^k_{i,j}, \quad -\sum_{i=1}^{L}\sum_{j=1}^{L} \pi^k_{i,j}, \quad [\sum_{i'=1}^{i}\sum_{j=1}^{L} \pi^k_{i',j}]_{i=1}^{L-1},$$
$$[-\sum_{i'=1}^{i}\sum_{j=1}^{L} \pi^k_{i',j}]_{i=1}^{L-1}, \quad [\sum_{j'=1}^{j}\sum_{i=1}^{L} \pi^k_{i,j'}]_{j=1}^{L-1}, \quad [-\sum_{j'=1}^{j}\sum_{i=1}^{L} \pi^k_{i,j'}]_{j=1}^{L-1}\Bigg]^T.$$

Let $f(b)$ denote the optimal value of the linear program as a function of b. Define:

$$b^* = \Bigg[0, \quad 1, \quad -1, \quad \left[F^k_C(i)\right]_{i=1}^{L-1}, \quad \left[-F^k_C(i)\right]_{i=1}^{L-1}, \quad \left[F^k_T(j)\right]_{j=1}^{L-1}, \quad \left[-F^k_T(j)\right]_{j=1}^{L-1}\Bigg]^T;$$

$$b^n = \Bigg[0, \quad 1, \quad -1, \quad \left[\hat{F}^k_C(i) + \overline{\epsilon}_k\right]_{i=1}^{L-1}, \quad \left[-\hat{F}^k_C(i) + \overline{\epsilon}_k\right]_{i=1}^{L-1},$$
$$\left[\hat{F}^k_T(j) + \overline{\epsilon}_k\right]_{j=1}^{L-1}, \quad \left[-\hat{F}^k_T(j) + \overline{\epsilon}_k\right]_{j=1}^{L-1}\Bigg]^T.$$

When $b = b^n$, LP is equivalent to the linear program for $\overline{\psi}^{\mathcal{R},X}_{l,k}$. Therefore, $f(b^n) = \overline{\psi}^{\mathcal{R},X}_{l,k} \leq 1$. When $b = b^*$, LP is equivalent to the linear program for $\psi^{\mathcal{R},X}_{l,k}$. Therefore, $f(b^*) = \psi^{\mathcal{R},X}_{l,k} \leq 1$. (LP is feasible when $b = b^*$ since P_0 is consistent with $\mathcal{R}$.) By Lemma 1 below, there exists a constant $C > 0$ such that $|f(b) - f(b^*)| \leq C\|b - b^*\|_2$ for any b with $f(b) < \infty$. This means that $|f(b^n) - f(b^*)| \leq C\|b^n - b^*\|_2$. Choose any $\eta > 0$. Then

$$\begin{aligned} P_0(|\overline{\psi}^{\mathcal{R},X}_{l,k} - \psi^{\mathcal{R},X}_{l,k}| > \eta) &= P_0(|f(b^n) - f(b^*)| > \eta) \\ &\leq P_0\left(\|b^n - b^*\|_2 > \frac{\eta}{C}\right). \end{aligned}$$

By Lemma 2 below, $\overline{\epsilon}_k \xrightarrow{p} 0$. *Also,* $\hat{F}_C^k(i) \xrightarrow{p} F_C^k(i)$ *and* $\hat{F}_T^k(j) \xrightarrow{p} F_T^k(j)$ *for all* $i, j = 1, .., L-1$ *by the Weak Law of Large Numbers. By Slutsky's Theorem,* $\hat{F}_C^k(i) + \overline{\epsilon}_k \xrightarrow{p} F_C^k(i), -\hat{F}_C^k(i) + \overline{\epsilon}_k \xrightarrow{p} -F_C^k(i), \hat{F}_T^k(j) + \overline{\epsilon}_k \xrightarrow{p} F_T^k(j)$, *and* $-\hat{F}_T^k(j) + \overline{\epsilon}_k \xrightarrow{p} -F_T^k(j)$ *for all* $i, j = 1, .., L-1$. *By Theorem 11.9 in Severini (2005),* $\lim_{n\to\infty} P_0\left(\|b^n - b^*\|_2 > \frac{\eta}{C}\right) = 0$. *Thus,*

$$\lim_{n\to\infty} P_0(|\overline{\psi}_{l,k}^{\mathcal{R},X} - \psi_{l,k}^{\mathcal{R},X}| > \eta) \;\; \leq \;\; \lim_{n\to\infty} P_0\left(\|b^n - b^*\|_2 > \frac{\eta}{C}\right) = 0,$$

so $\lim_{n\to\infty} P_0(|\overline{\psi}_{l,k}^{\mathcal{R},X} - \psi_{l,k}^{\mathcal{R},X}| > \eta) = 0$. *The choice of* η *was arbitrary, so we conclude that* $\overline{\psi}_{l,k}^{\mathcal{R},X} \xrightarrow{p} \psi_{l,k}^{\mathcal{R},X}$. *Analogously, it can be shown that* $\overline{\psi}_{u,k}^{\mathcal{R},X} \xrightarrow{p} \psi_{u,k}^{\mathcal{R},X}$. *Thus, (II) holds. We conclude that* $\overline{\psi}_l^{\mathcal{R},X}$ *and* $\overline{\psi}_u^{\mathcal{R},X}$ *are consistent estimators of* $\psi_l^{\mathcal{R},X}$ *and* $\psi_u^{\mathcal{R},X}$, *respectively.* □

Lemma 1 *Consider the linear programming problem:*

$$\min_x c^T x, \;\; \textit{subject to } Ax \leq b, x \geq 0, \tag{2.26}$$

where $A \in \mathbb{R}^{d_1 \times d_2}$. *Let* $f(b)$ *denote the optimal value of the linear program as a function of* b, *where we use the convention that* $f(b) = \infty$ *if the problem is infeasible. Consider any* b^* *such that* $f(b^*)$ *is finite (i.e., the linear program is bounded and feasible at* $b = b^*$*). Then there exists a constant* $C > 0$ *such that for any* b' *satisfying* $f(b') < \infty$, *we have*

$$|f(b') - f(b^*)| \leq C\|b' - b^*\|_2.$$

Proof *We assume the reader has familiarity with linear programming terminology. (For an overview of linear programming relevant to our proof, please see Chapter 6 of Dantzig and Thapa (2006).) Without loss of generality we can drop the* $x \geq 0$ *term in the linear program (2.26) since these constraints can be incorporated into the set of inequalities* $Ax \leq b$*. Consider the dual linear program:*

$$\max_y b^T y, \text{ subject to } A^T y = c, y \geq 0.$$

It is bounded and feasible at $b = b^*$ *and at* $b = b'$*, which follows from the conditions in the lemma. Let* V^* *denote the set of vertices of the dual linear program, which is non-empty, finite, and only depends on* A *and* c *(and does not depend on* b*). Since the optimal value of the dual problem occurs at a vertex, it equals* $\max\{y^T b : y \in V^*\}$*, for any vector* b *for which the linear program is bounded and feasible (which includes the cases* $b = b^*$ *and* $b = b'$*).*

By strong duality, for each $b \in \{b', b^*\}$*, the optimal value of the primal (original) linear program equals the optimal value of the dual linear program, and therefore*

$$|f(b') - f(b^*)| = \left| \max_{y \in V^*} y^T b' - \max_{y \in V^*} y^T b^* \right| \leq \max_{y \in V^*} |y^T (b' - b^*)| \leq \max_{y \in V^*} \|y\|_2 \|b' - b^*\|_2,$$

where the last inequality follows from the Cauchy-Schwartz inequality. The lemma is proved for $C = \max_{y \in V^*} \|y\|_2$*, which is finite since* V^* *is non-empty and has a finite number of elements.* □

Lemma 2 *Let P_0 be the true joint distribution on (X, Y_C, Y_T). Suppose Assumptions 1 and 2 hold. Then $\overline{\epsilon}_k \xrightarrow{p} 0$ for any k, as $n \to \infty$.*

Proof *In this proof, we will show that $\lim_{n\to\infty} P_0(|\overline{\epsilon}_k - 0| > \eta) = 0$ for any $\eta > 0$.*

With regards to notation used throughout this proof, limits are taken as $n \to \infty$. For example, $\xrightarrow{p}$ signifies convergence in probability as $n \to \infty$. Define 1_E to be a random variable which takes the value 1 if the event E occurs, and 0 otherwise. Generally, define $[z_{i,j}]_{i=1}^{L} = [z_{1,j}, z_{2,j}, ..., z_{L,j}]$ and $[z_{i,j}]_{j=1}^{L} = [z_{i,1}, z_{i,2}, ..., z_{i,L}]$. Also, define $\|y\|_2 = \sqrt{\sum_{i=1}^{M} y_i^2}$, where $y = (y_1, y_2, .., y_M)^T$.

Choose any $k = 1, .., K$. Fix $n \in \mathbb{N}$. Consider the following linear program, referred to as LP:

$$\min_x c^T x, \text{ subject to } Ax \leq b, x \geq 0.$$

Define x and c as:

$$\begin{aligned} x &= \left[[\pi_{i,1}^k]_{i=1}^{L}, \quad [\pi_{i,2}^k]_{i=1}^{L}, \quad \ldots, \quad [\pi_{i,L}^k]_{i=1}^{L}, \quad \epsilon_k \right]^T, \\ c^T &= \left[\mathbf{0}_{1\times L^2}, \quad 1 \right], \end{aligned}$$

where $\mathbf{0}_{1\times L^2}$ is a row vector of length L^2 containing only zeroes. Define A as the matrix that satisfies:

$$Ax = \Big[\textstyle\sum_{g(i,j)=0} \pi_{i,j}^k, \quad \sum_{i=1}^{L}\sum_{j=1}^{L} \pi_{i,j}^k, \quad -\sum_{i=1}^{L}\sum_{j=1}^{L} \pi_{i,j}^k, \quad [\sum_{i'=1}^{i}\sum_{j=1}^{L} \pi_{i',j}^k - \epsilon_k]_{i=1}^{L-1},$$
$$[-\textstyle\sum_{i'=1}^{i}\sum_{j=1}^{L} \pi_{i',j}^k - \epsilon_k]_{i=1}^{L-1}, \quad [\sum_{j'=1}^{j}\sum_{i=1}^{L} \pi_{i,j'}^k - \epsilon_k]_{j=1}^{L-1}, \quad [-\textstyle\sum_{j'=1}^{j}\sum_{i=1}^{L} \pi_{i,j'}^k - \epsilon_k]_{j=1}^{L-1} \Big]^T.$$

Let $f(b)$ be the optimal value of LP as a function of b. Define:

$$b^* = \left[0,\ \ 1,\ \ -1,\ \ \left[F_C^k(i)\right]_{i=1}^{L-1},\ \ \left[-F_C^k(i)\right]_{i=1}^{L-1},\ \ \left[F_T^k(j)\right]_{j=1}^{L-1},\ \ \left[-F_T^k(j)\right]_{j=1}^{L-1}\right]^T,$$

$$b^n = \left[0,\ \ 1,\ \ -1,\ \ \left[\hat{F}_C^k(i)\right]_{i=1}^{L-1},\ \ \left[-\hat{F}_C^k(i)\right]_{i=1}^{L-1},\ \ \left[\hat{F}_T^k(j)\right]_{j=1}^{L-1},\ \ \left[-\hat{F}_T^k(j)\right]_{j=1}^{L-1}\right]^T.$$

By Assumption 1, $f(b^*) = 0$. When $b = b^n$, LP is equivalent to the linear program for $\overline{\epsilon}_k$, so $f(b^n) = \overline{\epsilon}_k \leq 1$. By Lemma 1, there exists a constant $C > 0$ such that $|f(b) - f(b^*)| \leq C\|b - b^*\|_2$ for any b such that $f(b) < \infty$. This means that $|f(b^n) - f(b^*)| \leq C\|b^n - b^*\|_2$.

Choose any $\eta > 0$. Then

$$P_0(|\overline{\epsilon}_k - 0| > \eta) = P_0(|f(b^n) - f(b^*)| > \eta) \leq P_0\left(\|b^n - b^*\|_2 > \frac{\eta}{C}\right).$$

It follows from the Weak Law of Large Numbers that b^n converges to b^* in probability, and so the right side of the above display converges to 0 as $n \to \infty$, completing the proof of the lemma.

□

Theorems 1 and 3 imply that, if P_0 is consistent with $\mathcal{R}$, then the probability limits of the estimators $\overline{\psi}_l^{\mathcal{R},X}, \overline{\psi}_l^{\mathcal{R}}, \overline{\psi}_l^{X}, \overline{\psi}_l$ satisfy the inequalities in Theorem 1. This means that including a baseline variable or restriction can only improve (or leave unchanged) the limiting value of the bound estimators. However, at a given sample

size, neither the plug-in estimators from Section 2.4.1 nor the above estimators are guaranteed to satisfy the corresponding inequalities in Theorem 1.

2.4.4 Non-regularness of the Proposed Estimator

Our estimator can be non-regular when the parameter (representing the lower or upper bound) is 0 or 1, which also occurs for Gadbury et al. (2004); Fan and Park (2009, 2010). Furthermore, our estimator can be non-regular at boundary cases as defined in the last paragraph of Section 2.4.1. Intuitively, non-regularity means that the asymptotic distribution of the estimator can change dramatically under small perturbations of the data generating distribution. Formally, an estimator $\widehat{\beta}_n$ of β is non-regular if, for some sequence of distributions $P^{(n)}$ satisfying $\sqrt{n}||P^{(n)} - P_0|| = O(1)$, the distribution of $\sqrt{n}(\widehat{\beta}_n - \beta(P^{(n)}))$ under $P^{(n)}$ converges to a different limit than $\sqrt{n}(\widehat{\beta}_n - \beta(P_0))$ under P_0, where $||\cdot||$ is total variation distance (Durrett, 2010).

To show an example of non-regularity for our problem, consider a binary outcome. Let P_0 be the unique joint distribution on (Y_C, Y_T) satisfying: $P_0(Y_C = y, Y_T = y) = 0.5$ for each $y \in \{1, 2\}$, and $P_0(Y_C \neq Y_T) = 0$. The lower bound $\psi_l(P_0)$ is 0. Let the randomization probability θ be 0.5. Under P_0, the distribution of $\sqrt{n}(\overline{\psi}_l - \psi_l(P_0))$ converges to $\max\{N(0, 1), 0\}$. The proof is presented at the end of this section. Let $P^{(n)}$ be the joint distribution with: $P^{(n)}(Y_C = 1, Y_T = 1) = 0.5$, $P^{(n)}(Y_C = 1, Y_T = 2) = 1/\sqrt{n}$, $P^{(n)}(Y_C = 2, Y_T = 1) = 0$, $P^{(n)}(Y_C = 2, Y_T = 2) = 0.5 - 1/\sqrt{n}$. We have $\sqrt{n}||P^{(n)} - P_0||$ is $O(1)$.

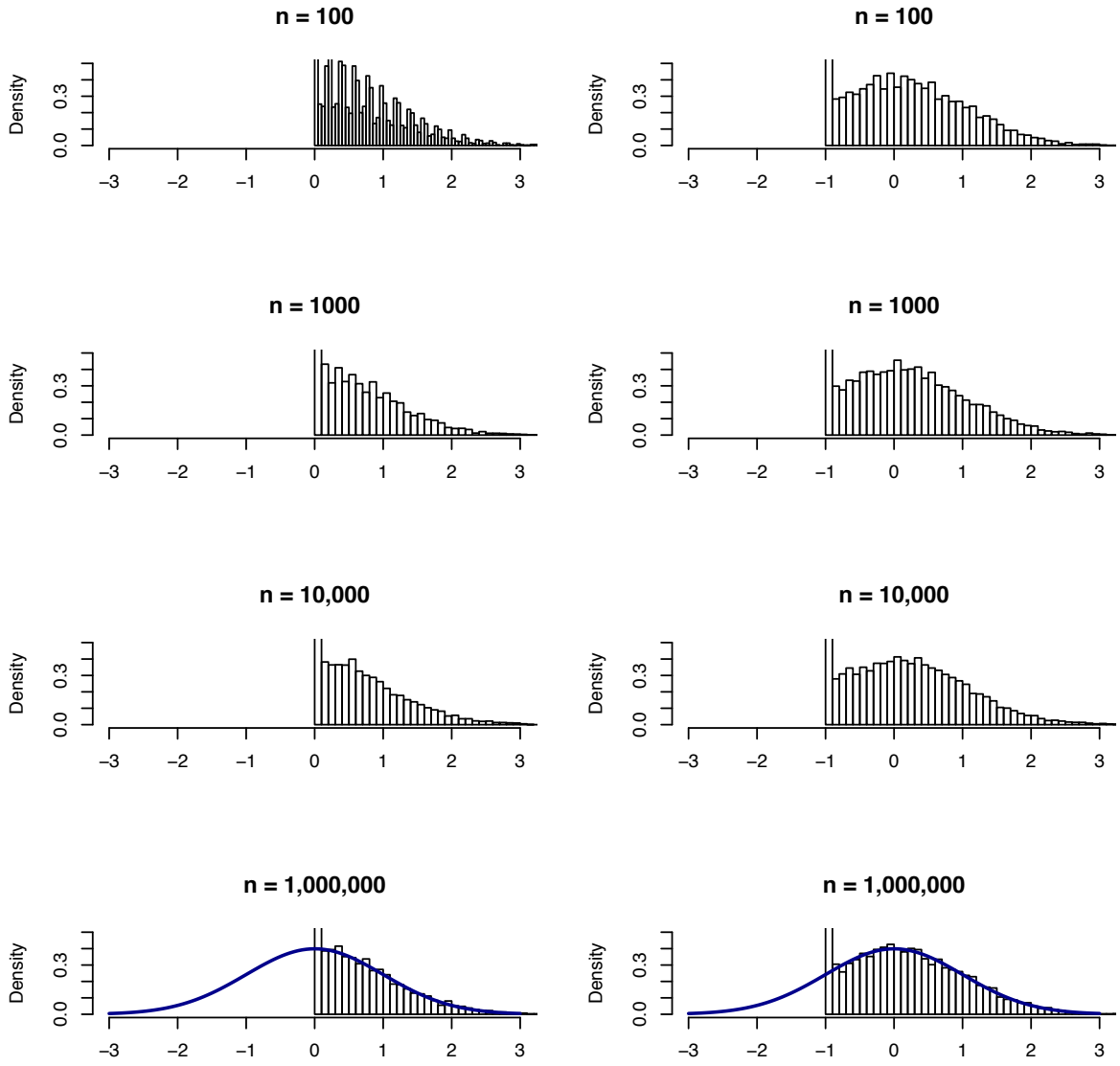


Figure 2.3: Distributions of $\sqrt{n}(\overline{\psi}_l - \psi_l(P_0))$ under P_0 (left panel), and of $\sqrt{n}(\overline{\psi}_l - \psi_l(P^{(n)}))$ under $P^{(n)}$ (right panel), for $n = 100, 1000, 10000$, and 1000000.

Proof *Consider any positive integer n. By Section 3.6 in Durrett (2010),*

$$\begin{aligned} ||P^{(n)} - P_0|| = & \quad \tfrac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2}\left|P^{(n)}(Y_C = i, Y_T = j) - P_0(Y_C = i, Y_T = j)\right| \\ = & \quad \tfrac{1}{2}\left[|0.5 - 0.5| + \left|\tfrac{1}{\sqrt{n}} - 0\right| + |0 - 0| + \left|0.5 - \tfrac{1}{\sqrt{n}} - 0.5\right|\right] = \tfrac{1}{\sqrt{n}}. \end{aligned}$$

Therefore, we have $\sqrt{n}||P^{(n)} - P_0|| = 1$. *Since n is an arbitrary positive integer, we conclude that* $\sqrt{n}||P^{(n)} - P_0|| \leq 1$ *for all integers* $n \geq 1$. *It follows that* $\sqrt{n}||P^{(n)} - P_0||$ *is* $O(1)$. □

Under $P^{(n)}$, the distribution of $\sqrt{n}(\overline{\psi}_l - \psi_l(P^{(n)}))$ converges to $\max\{N(0,1), -1\}$, not $\max\{N(0,1), 0\}$. The proof is shown at the end of this section. Intuitively, the sequence $P^{(n)}$ is like P_0 except it makes the small perturbation $1/\sqrt{n}$ to the marginal distribution under control, which results in a strikingly different limit distribution than under P_0. Figure 2.3 illustrates the above behavior using simulations. It shows the distributions of $\sqrt{n}(\overline{\psi}_l - \psi_l(P_0))$ under P_0 (left panel), and of $\sqrt{n}(\overline{\psi}_l - \psi_l(P^{(n)}))$ under $P^{(n)}$ (right panel), for $n = 100, 1000, 10000$, and 1000000. The results are consistent with the theoretical results that the distribution in the left panel converges to $\max\{N(0,1), 0\}$ and the distribution in the right panel converges to $\max\{N(0,1), -1\}$. The $N(0,1)$ density curve is superimposed on the two plots where $n = 1000000$. The point mass at 0 (-1) in the left (right) plot is approximately 0.49 (0.16). In this example, the above limit distributions are the same if we modify the parameter and estimator to incorporate the no harm assumption.

Below is the proof of the limiting distribution of $\sqrt{n}(\overline{\psi}_l - \psi_l(P_0))$ under P_0, followed by the proof of the limiting distribution of $\sqrt{n}(\overline{\psi}_l - \psi_l(P^{(n)}))$ under $P^{(n)}$.

Proof (Limiting distribution of $\sqrt{n}(\overline{\psi}_l - \psi_l(P_0))$, under P_0) *We prove that, under P_0, $\sqrt{n}(\overline{\psi}_l - \psi_l(P_0))$ converges in distribution to* $\max\{N(0,1), 0\}$.

Let n denote the number of participants in the randomized trial. For any given participant m, let

$$Z_m = \begin{pmatrix} 1(A_m = 0, Y_m = 2) & 1(A_m = 1, Y_m = 2) & 1(A_m = 1) \end{pmatrix}^T.$$

Let $\beta = \begin{pmatrix} \beta_C & \beta_T & \theta \end{pmatrix}^T = E_{P_0}[Z_m] = \begin{pmatrix} 0.25 & 0.25 & 0.5 \end{pmatrix}^T$.

By the Multivariate Central Limit Theorem (van der Vaart, 2000),

$$\sqrt{n}\left(\frac{1}{n}\sum_{m=1}^{n} Z_m - E_{P_0}[Z_1]\right) = \sqrt{n}\left(\begin{pmatrix} \frac{1}{n}\sum_{m=1}^{n} 1(A_m = 0, Y_m = 2) \\ \frac{1}{n}\sum_{m=1}^{n} 1(A_m = 1, Y_m = 2) \\ \frac{1}{n}\sum_{m=1}^{n} 1(A_m = 1) \end{pmatrix} - \begin{pmatrix} \beta_C \\ \beta_T \\ \theta \end{pmatrix}\right)$$

converges in distribution to $N(0, \Sigma)$, where

$$\Sigma = \begin{pmatrix} \beta_C(1-\beta_C) & -\beta_C\beta_T & -\beta_C\theta \\ -\beta_C\beta_T & \beta_T(1-\beta_T) & \beta_T(1-\theta) \\ -\beta_C\theta & \beta_T(1-\theta) & \theta(1-\theta) \end{pmatrix} = \begin{pmatrix} 0.1875 & -0.0625 & -0.125 \\ -0.0625 & 0.1875 & 0.125 \\ -0.125 & 0.125 & 0.25 \end{pmatrix}.$$

Define $g((x, y, z)^T) = y/z - x/(1-z)$. By the Multivariate Delta Method (Rohde,

2014),

$$\sqrt{n}\left(g\left(\frac{1}{n}\sum_{m=1}^{n}Z_m\right)-g\left(E_{P_0}[Z_1]\right)\right)\xrightarrow{d}$$
$$N\left(0,\frac{\beta_C(1-\beta_C)}{(1-\theta)^2}+\frac{\beta_T(1-\beta_T)}{\theta^2}-\frac{\beta_C^2\theta}{(1-\theta)^3}-\frac{\beta_T^2(1-\theta)}{\theta^3}\right).$$

After simplifying both sides, we have

$$\sqrt{n}\left(\hat{p}_T(2)-\hat{p}_C(2)\right)\xrightarrow{d}N(0,1).$$

By applying the Continuous Mapping Theorem (van der Vaart, 2000) with $h(x)=\max\{x,0\}$*, we have* $\max\{\sqrt{n}\left(\hat{p}_T(2)-\hat{p}_C(2)\right),0\}\xrightarrow{d}\max\{N(0,1),0\}$*. Also, we have* $\max\{\sqrt{n}\left(\hat{p}_T(2)-\hat{p}_C(2)\right),0\}=\sqrt{n}\max\{\hat{p}_T(2)-\hat{p}_C(2),0\}$*. Thus,*

$$\sqrt{n}\max\{\hat{p}_T(2)-\hat{p}_C(2),0\}\xrightarrow{d}\max\{N(0,1),0\}.$$

Since $\psi_l(P_0)=0$ *and* $\overline{\psi}_l=\max\{\hat{p}_T(2)-\hat{p}_C(2),0\}$ *(Gadbury et al., 2004), we have*

$$\sqrt{n}(\overline{\psi}_l-\psi_l(P_0))\xrightarrow{d}\max\{N(0,1),0\}.$$

□

Proof (Limiting distribution of $\sqrt{n}(\overline{\psi}_l-\psi_l(P^{(n)}))$, under $P^{(n)}$) *We prove that, under* $P^{(n)}$*,* $\sqrt{n}(\overline{\psi}_l-\psi_l(P^{(n)}))$ *converges in distribution to* $\max\{N(0,1),-1\}$*.*

Let n denote the number of participants in the randomized trial. For any given participant m, let

$$Z_{n,m} = \frac{1}{\sqrt{n}} \left(1(A_m = 0, Y_m = 2) + \tfrac{0.5}{\sqrt{n}} \quad 1(A_m = 1, Y_m = 2) \quad 1(A_m = 1) \right)^T .$$

Let $\beta_n = \frac{1}{\sqrt{n}} \left(\beta_{C,n} \quad \beta_{T,n} \quad \theta_n \right)^T = E_{P_n}[Z_{n,m}] = \frac{1}{\sqrt{n}} \left(0.25 \quad 0.25 \quad 0.5 \right)^T .$

By the Lindeberg-Feller Central Limit Theorem (van der Vaart, 2000),

$$\sum_{m=1}^{n} \left(Z_{n,m} - E_{P^{(n)}}[Z_{n,m}] \right) \xrightarrow{d} N(0, \Sigma), \text{ where}$$

$$\Sigma = \lim_{n\to\infty} \left(\sum_{m=1}^{n} Cov\ Z_{n,m} \right) = \lim_{n\to\infty} \begin{pmatrix} 0.1875 - \frac{0.5}{n} & -0.0625 + \frac{0.125}{\sqrt{n}} & -0.125 + \frac{0.25}{\sqrt{n}} \\ -0.0625 & 0.1875 & 0.125 \\ -0.125 & 0.125 & 0.25 \end{pmatrix}$$

$$= \begin{pmatrix} 0.1875 & -0.0625 & -0.125 \\ -0.0625 & 0.1875 & 0.125 \\ -0.125 & 0.125 & 0.25 \end{pmatrix} .$$

It can be shown that the condition $\lim_{n\to\infty} \sum_{m=1}^{n} E\left[||Z_{n,m}||^2 1(||Z_{n,m}|| > \epsilon) \right] = 0$ *is*

satisfied for every $\epsilon > 0$. It follows that,

$$\sqrt{n}\left(\begin{pmatrix} \frac{1}{n}\sum_{m=1}^{n} 1(A_m = 0, Y_m = 2) + \frac{0.5}{\sqrt{n}} \\ \frac{1}{n}\sum_{m=1}^{n} 1(A_m = 1, Y_m = 2) \\ \frac{1}{n}\sum_{m=1}^{n} 1(A_m = 1) \end{pmatrix} - \begin{pmatrix} 0.25 \\ 0.25 \\ 0.5 \end{pmatrix}\right) \xrightarrow{d} N(0, \Sigma).$$

Define $g((x, y, z)^T) = y/z - x/(1-z)$. By the Multivariate Delta Method (Rohde, 2014),

$$\sqrt{n}\left(g\left(\begin{pmatrix} \frac{1}{n}\sum_{m=1}^{n} 1(A_m = 0, Y_m = 2) + \frac{0.5}{\sqrt{n}} \\ \frac{1}{n}\sum_{m=1}^{n} 1(A_m = 1, Y_m = 2) \\ \frac{1}{n}\sum_{m=1}^{n} 1(A_m = 1) \end{pmatrix}\right) - g\left(\begin{pmatrix} 0.25 \\ 0.25 \\ 0.5 \end{pmatrix}\right)\right) \xrightarrow{d} N\left(0, 1\right).$$

After simplifying, we have

$$\sqrt{n}\left(\hat{p}_T(2) - \hat{p}_C(2) - \frac{1}{\sqrt{n}}\right) + \left(1 - \frac{0.5}{1 - \frac{1}{n}\sum_{m=1}^{n} 1(A_m = 1)}\right) \xrightarrow{d} N(0, 1).$$

By Slutsky's Lemma (van der Vaart, 2000),

$$\sqrt{n}\left(\hat{p}_T(2) - \hat{p}_C(2) - \frac{1}{\sqrt{n}}\right) \xrightarrow{d} N(0, 1).$$

Applying the Continuous Mapping Theorem (van der Vaart, 2000) with $h(x) = \max\{x, -1\}$, we have $\max\left\{\sqrt{n}\left(\hat{p}_T(2) - \hat{p}_C(2) - \frac{1}{\sqrt{n}}\right), -1\right\} \xrightarrow{d} \max\{N(0, 1), -1\}$. This simplifies

to:

$$\sqrt{n}\left[\max\{\hat{p}_T(2)-\hat{p}_C(2),0\}-\frac{1}{\sqrt{n}}\right] \xrightarrow{d} \max\{N(0,1),-1\}.$$

Since $\psi_l(P^{(n)}) = \max\{P_n(Y_T=2)-P_n(Y_C=2),0\} = \frac{1}{\sqrt{n}}$ *and* $\overline{\psi}_l = \max\{\hat{p}_T(2)-\hat{p}_C(2),0\}$ *(Gadbury et al., 2004), we have*

$$\sqrt{n}(\overline{\psi}_l - \psi_l(P^{(n)})) \xrightarrow{d} \max\{N(0,1),-1\}.$$

□

2.4.5 Inference Based on the Proposed Estimator

The impact of non-regularity is that confidence intervals based on the standard nonparametric bootstrap (called the n-bootstrap) are typically inconsistent, as shown by Bickel et al. (1997). They recommend to remedy this by using the m-out-of-n bootstrap, where each bootstrap replicate data set is generated by resampling $m \leq n$ participants with replacement. Fan and Park (2010) use m-out-of-n bootstrap to construct confidence intervals, and report that coverage probability is relatively close to the desired 95% level in their simulations. We also use the m-out-of-n bootstrap in our simulations in Section 2.6. Just as Fan and Park (2010), we select m based on Bickel and Sakov (2008), whose algorithm aims to achieve pointwise asymptotic consistency and efficiency. Consistency requires that in cases where the $n-$bootstrap is inconsistent (as can happen when the estimator is non-regular, for example), that

the growth rate of m be slow enough (e.g., $m/n \to 0$ and $m \to \infty$). Efficiency requires that when the $n-$bootstrap is consistent, that the growth rate of m be of similar order as n. The procedure that we apply to choose m is as follows:

1. The candidates for m are taken to be $m_j = [0.95^j n]$, $j = 0, 1, 2, ...$, that satisfy $m_j \geq 20$. The function $[\cdot]$ is the ceiling function.

2. For any given candidate m_j, we generate 5000 bootstrap data sets each by resampling m_j participants, with replacement, from the data set. The lower bound estimate is computed for each bootstrap data set. If any of the 5000 lower bound estimates are undefined, m_j is eliminated from consideration; an undefined bound estimate indicates that there were no control or no treatment participants in the bootstrap data set, and this suggests the candidate m_j is too small. Otherwise, let the function $Q^*_{m_j}$ be the vector of 5000 lower bound estimates from smallest to largest.

3. For any given m_j, let m'_j denote the candidate that is the next smallest after m_j and is still under consideration. The value m is chosen to be $\text{argmin}_{m_j} \rho(Q^*_{m_j}, Q^*_{m'_j})$, where the function ρ takes the absolute value of the difference between the two vectors and returns the maximum element. This value m is used to construct the CI for the lower bound. If multiple m_j achieve the minimum, the largest value is chosen. The argmin is taken over only m_j that are under consideration. An analogous procedure is used to choose the value m for constructing the CI

for the upper bound.

For a given data generating distribution, the m-out-of-n bootstrap has asymptotically correct coverage probability (called pointwise consistency) for our problem if both $m \to \infty$ and $m/n \to 0$ as $n \to \infty$; this result follows from Theorem 1 of Bickel et al. (1997). The convergence rate of m is difficult to determine, due to the complexity of the Bickel and Sakov (2008) algorithm. It is difficult even to establish pointwise consistency; the proof of this property in Bickel and Sakov (2008) requires six assumptions that would be very hard to verify from data. Therefore, just as for Fan and Park (2010), the resulting confidence intervals may fail to be pointwise consistent. Fan and Park (2009, Section 5.2) give an alternative approach requiring substantially weaker (but still hard to verify) assumptions. An important problem addressed in Chapter 3 is to construct confidence intervals that overcome the above issues. Despite the lack of asymptotic guarantees, the m-out-of-n bootstrap has relatively good performance in our simulation studies at sample size 500 or greater.

Our estimator can have substantial bias (in terms of its contribution to the mean squared error) in finite samples (Section 2.6), just as the estimators of Fan and Park (2009, 2010). They derive a first-order bias correction for their estimator. In our case, deriving a general bias correction would be quite challenging since our estimator does not have a simple analytic form (and instead is represented as a solution to linear programs).

Define the asymptotic distribution of our estimator $\overline{\psi}_l^{\mathcal{R}}$ as the limit of $\sqrt{n}(\overline{\psi}_l^{\mathcal{R}} - \psi_l^{\mathcal{R}})$

as $n \to \infty$ under P_0. If this is not a boundary case (as defined in Section 2.4.1), then the asymptotic distribution is the maximum of the components of a mean zero (possibly degenerate) multivariate normal distribution with covariance matrix depending on P_0 and $\mathcal{R}$. For a boundary case, the asymptotic distribution can be more complex since then $\sqrt{n\bar{\epsilon}}$ has a non-degenerate limit distribution and affects the asymptotic distribution of our estimator. It is an open problem to precisely characterize the limit distribution in boundary cases; however, even if this were solved, it would not immediately lead to a confidence interval procedure since the limit distribution would generally depend on the unknown P_0.

2.5 MISTIE Application using Bound Estimator from Section 2.4.2

Using MISTIE II, we estimate bounds on the fraction of ICH patients who benefit from treatment relative to control. We apply the estimators from Section 2.4.2.

2.5.1 30- and 180-day mRS scores

For both 30- and 180-day mRS, four types of sharp lower/upper bounds are estimated: (i) (ψ_l, ψ_u), (ii) $(\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}})$, (iii) (ψ_l^{X}, ψ_u^{X}), (iv) $(\psi_l^{\mathcal{R},X}, \psi_u^{\mathcal{R},X})$. The restrictions $\mathcal{R}$ considered are Benefit $\leq d$ levels and Harm $\leq d$ levels. The value d is varied from 1

to 5 for the former, and 0 to 5 for the latter. The baseline variable X is stroke severity as measured by the National Institutes of Health Stroke Scale (NIHSS), where a stroke is classified as non-severe if the score ≤ 20 and severe otherwise (Kreutzer et al., 2011). When estimating bounds for a given outcome (e.g., 180-day mRS), we exclude participants who are missing that outcome; for both mRS outcomes, we exclude the single patient with missing baseline NIHSS score. The resulting sample sizes are 53 treatment and 39 control participants for 30-day mRS, and 52 treatment and 37 control participants for 180-day mRS. The top panel of Figure 2.4 shows the empirical distributions under treatment and control of 30-day mRS. The middle panel shows the empirical distributions within the non-severe stratum, and the bottom panel shows the empirical distributions within the severe stratum. Figure 2.5 for the 180-day mRS outcome is organized analogously. The proportion in each subpopulation is estimated by the corresponding sample proportion of MISTIE II participants after excluding participants as described above.

The bound estimates for 30-day mRS and 180-day mRS are plotted in Figures 2.6 and 2.7, respectively. Each bar ranges from the lower to the upper bound estimate. A bar is grey if the baseline variable is not used, and black otherwise. The restriction imposed, if any, is indicated on the x-axis. For conciseness, restrictions whose grey and black bars are identical to those under no restrictions are excluded from these figures. For grey bars, the value of $\bar{\epsilon}$ (defined in Section 2.4.2) is listed above the bar, if it is nonzero. For black bars, ϵ=* indicates that one or more of the $\bar{\epsilon}_k$'s is nonzero.

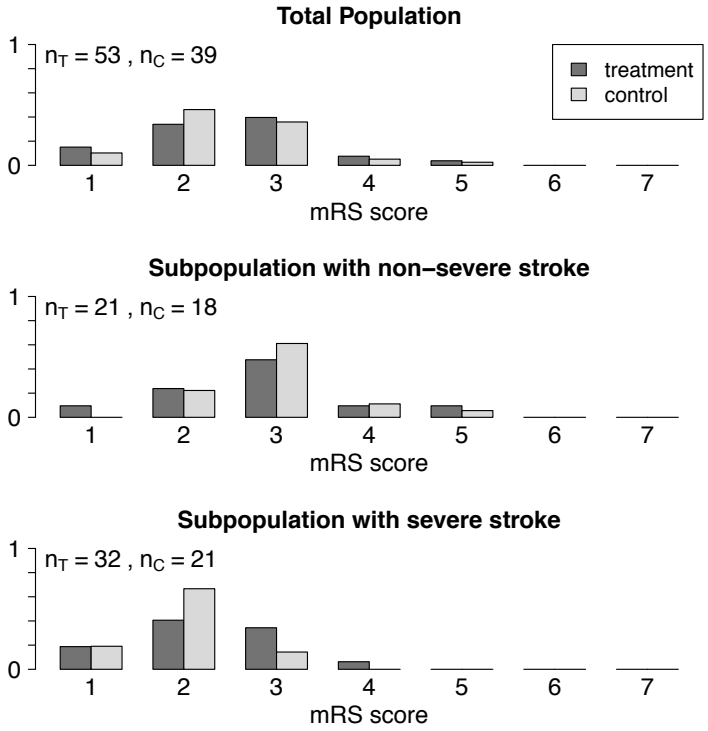


Figure 2.4: Empirical Distribution of 30-day mRS Score under Treatment and Control.

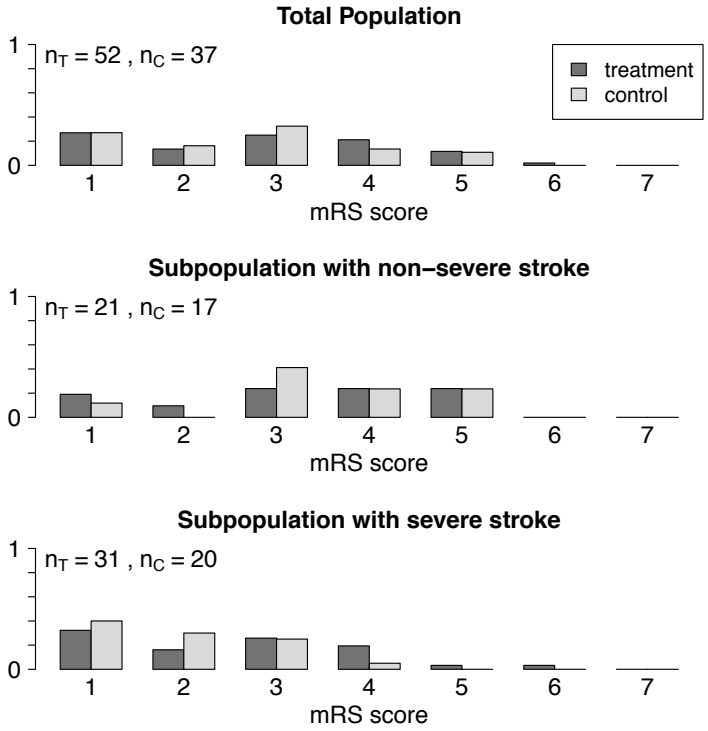


Figure 2.5: Empirical Distribution of 180-day mRS Score under Treatment and Control.

The values of the bound estimates are recorded in Tables 5.1 and 5.2 of Chapter 5.

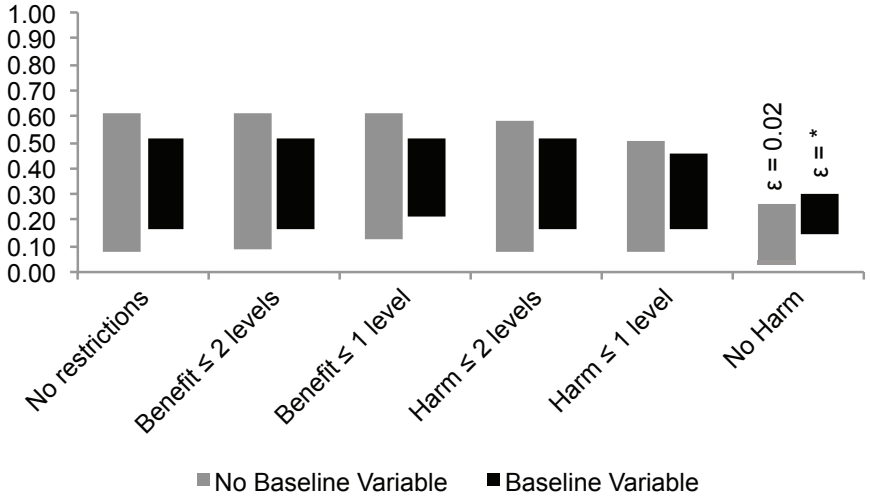


Figure 2.6: Bounds Estimates for 30-day mRS Score, using Method from Section 2.4.2.

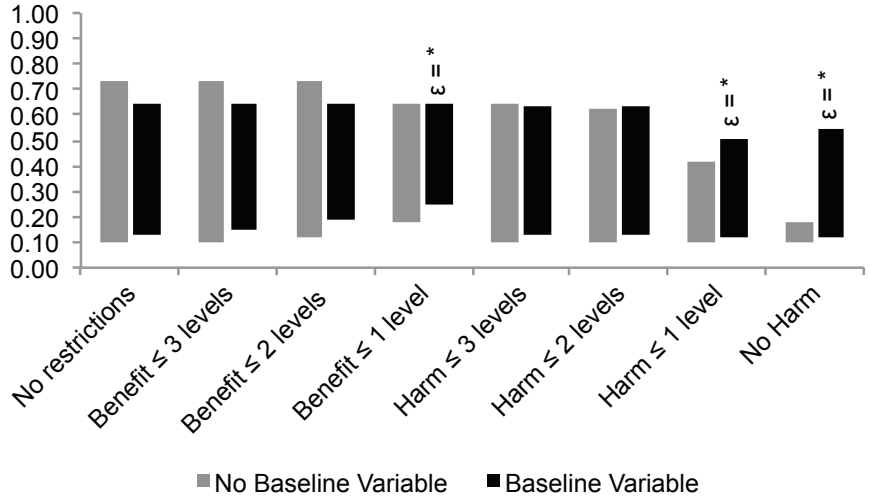


Figure 2.7: Bounds Estimates for 180-day mRS Score, using Method from Section 2.4.2.

The pair of estimated bounds $\left[\overline{\psi}_l, \overline{\psi}_u\right]$ is [0.07,0.61] for 30-day mRS, and [0.10,0.73] for 180-day mRS. The widths of these estimated bounds, i.e., the difference between

the upper and lower bound estimates, are 0.54 and 0.63, respectively. Restrictions and the baseline variable can narrow the width of the estimated bounds. For 180-day mRS, the width narrows by 0.17 under Benefit ≤ 1, 0.31 under Harm ≤ 1, and 0.55 under no harm, relative to no restrictions. These reductions are absolute differences in widths, as is the case throughout this chapter. Without restrictions, the baseline variable narrows the width by 0.19 for 30-day mRS, and 0.12 for 180-day mRS. With the restriction $\mathcal{R} = \{\text{Harm } \leq 2\}$, the upper bound estimate with the baseline variable ($\overline{\psi}_u^{\mathcal{R},X} = 0.63$) is slightly above that without the baseline variable ($\overline{\psi}_u^{\mathcal{R}} = 0.62$). This can occur since, as mentioned in Section 2.4.2, the bound estimators need not obey the corresponding inequalities in Theorem 1.

In Figures 2.6 and 2.7, there are five cases in which $\bar{\epsilon} > 0$ or some $\bar{\epsilon}_k > 0$. (The value $\bar{\epsilon}$ is the analog of $\bar{\epsilon}_k$ when no baseline variable is used. It was defined in Section 2.4.2.) We point out two features. First, these bound estimates may not be contained within the interval formed by estimates under a less stringent restriction. For 30-day mRS, the lower bound estimate under $\mathcal{R} = \{\text{no harm}\}$ is $\overline{\psi}_l^{\mathcal{R}} = 0.03$, which is below the lower bound estimate $\overline{\psi}_l^{\mathcal{R}'} = 0.07$ under the weaker restriction $\mathcal{R}' = \{\text{Harm } \leq 1\}$. This behavior is either due to a boundary case (see Section 4), small sample performance of the estimator, or the data generating distribution not satisfying the no harm assumption. In the third case, the bound estimators may be inconsistent. Second, for a given restriction, an upper bound estimate can be much larger, or a lower bound estimate much smaller, with the baseline variable

than without it. For 180-day mRS, under the restriction $\mathcal{R} = \{\text{no harm}\}$, the upper bound estimate is $\overline{\psi}_u^{\mathcal{R}} = 0.18$ without the baseline variable, and $\overline{\psi}_u^{\mathcal{R},X} = 0.54$ with it. One possible cause for this behavior is that the no harm assumption is false. In this case, the parameter $\psi_u^{\mathcal{R}}$ could be well-defined while $\psi_u^{\mathcal{R},X}$ is undefined, as discussed in Section 2.3.2. Then $\overline{\psi}_u^{\mathcal{R}}$ could be much smaller than $\overline{\psi}_u^{\mathcal{R},X}$, even at large sample sizes.

2.5.2 Reduction in clot volume

Reduction in clot volume (RICV) is the difference between clot volume at baseline and end of treatment, as defined by Mould et al. (2013). In MISTIE II, the observed RICV range was [-2.57, 75.45] mL under treatment, and [-14.86, 12.01] mL under control. For any given baseline clot volume, the larger the RICV, the better.

We discretize RICV to an ordinal outcome. The appropriate bin length depends on the change in RICV that would be a clinically meaningful difference. If an improvement of 5 mL or more is a clinically meaningful benefit, the bin length can be set to 5 mL. Based on personal communications with neurologist Daniel Hanley, there currently is not enough biologic evidence to define a clinically meaningful change. Therefore, we consider various bin lengths, including 2, 5, 10, and 20 mL. We name the corresponding ordinal outcomes RICV2, RICV5, RICV10, and RICV20, respectively. Let y be (continuous) reduction in clot volume. We define RICV2, RICV5, RICV10, and RICV20 as follows:

Definition of RICV2

- Level 1: $y < 0$ mL
- Level 2: 0 mL $\leq y < 2$ mL
- Level 3: 2 mL $\leq y < 4$ mL
- Level 4: 4 mL $\leq y < 6$ mL
- Level 5: 6 mL $\leq y < 8$ mL
- Level 6: 8 mL $\leq y < 10$ mL
- Level 7: 10 mL $\leq y < 12$ mL
- Level 8: 12 mL $\leq y < 14$ mL
- Level 9: 14 mL $\leq y < 16$ mL
- Level 10: 16 mL $\leq y < 18$ mL
- Level 11: 18 mL $\leq y < 20$ mL
- Level 12: $y \geq 20$ mL

Definition of RICV5

- Level 1: $y < 0$ mL
- Level 2: 0 mL $\leq y < 5$ mL
- Level 3: 5 mL $\leq y < 10$ mL

- Level 4: 10 mL $\leq y <$ 15 mL
- Level 5: 15 mL $\leq y <$ 20 mL
- Level 6: $y \geq$ 20 mL

Definition of RICV10

- Level 1: $y <$ 0 mL
- Level 2: 0 mL $\leq y <$ 10 mL
- Level 3: 10 mL $\leq y <$ 20 mL
- Level 4: $y \geq$ 20 mL

Definition of RICV20

- Level 1: $y <$ 0 mL
- Level 2: 0 mL $\leq y <$ 20 mL
- Level 3: $y \geq$ 20 mL

We now present the procedure and results of the RICV5 analysis. The analysis procedure is analogous for RICV2, RICV10, and RICV20. The fraction who benefit, with respect to RICV5, is the fraction who would have a higher RICV5 under treatment than under control. We estimate sharp bounds (i)-(iv) as in Section 2.5.1. The restrictions $\mathcal{R}$ considered are Benefit $\leq d$ levels and Harm $\leq d$ levels, where d is

varied from 1 to 4 for the former, and 0 to 4 for the latter. The baseline variable X is an indicator of the baseline clot volume being above or below the median baseline clot volume of the MISTIE II participants (43.2 mL). There are no missing data, and all MISTIE II participants are included in the analysis. The empirical distributions of RICV5 under treatment and control, with and without stratifying by baseline clot volume, are shown in Figure 2.8. While all control participants had RICV5 of 4 or less, 74% of treatment participants had RICV5 higher than 4. This suggests that treatment has a major effect on RICV5.

The estimated bounds on the fraction who benefit are plotted in Figure 2.9. The values are recorded in Table 5.3 of Chapter 5. The estimated bounds are [0.82, 0.96] with neither the baseline variable nor restrictions, and [0.83, 0.96] with only the baseline variable. Assuming Benefit $\leq d$ levels (d = 1,2, or 3), the bound estimates are much wider than without restrictions. The values $\overline{\epsilon}$ and $\overline{\epsilon}_k$ range from 0.12 to 0.43. Large values of $\overline{\epsilon}$ or $\overline{\epsilon}_k$ raise doubts about the validity of the restrictions; it is an area of future work to construct formal hypothesis tests, to determine with high confidence whether a large observed value of $\overline{\epsilon}$ or $\overline{\epsilon}_k$ can be explained by chance variation or is due to violations of the restrictions. The restrictions on harm are not shown because the results are the same as under no restrictions.

The results for RICV2, RICV10, and RICV20 are shown in Tables 5.4, 5.5, 5.6 and Figures 5.1, 5.2, 5.3 of Chapter 5. The bound estimates are almost identical among RICV2, RICV5, and RICV10. The estimates for RICV20 are smaller because many

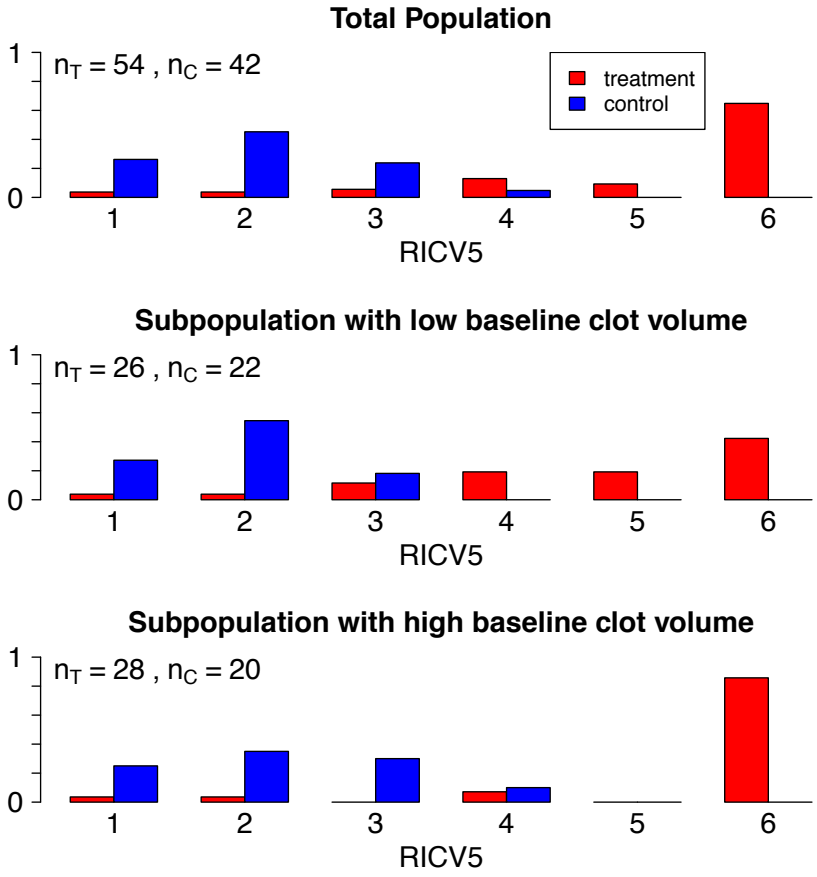


Figure 2.8: Empirical Probability Mass Functions of RICV5 under Treatment and Control.

improvements that would be benefits at the smaller bin lengths no longer qualify when the bin length is 20 mL.

Using m-out-of-n bootstrap, we compute two-sided 95% CI's for the lower bound ψ_l and the upper bound ψ_u for all outcomes. To construct these CI's, we follow the procedure previously presented in Section 2.4.5 to choose m and apply the percentile method. The CI's for ψ_l and ψ_u are $[0, 0.29]$ and $[0.42, 0.76]$ for 30-day mRS; $[0, 0.31]$ and $[0.50, 0.85]$ for 180-day mRS; $[0.71, 0.92]$ and $[0.90, 1]$ for RICV5. These CI's

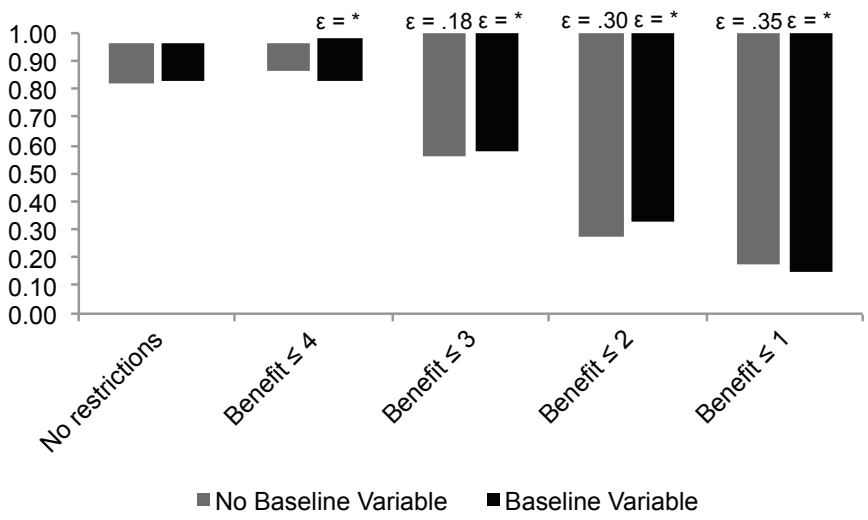


Figure 2.9: Bound Estimates for RICV5, using Method from Section 2.4.2.

should be interpreted with caution. The m-out-of-n bootstrap can have lower than nominal coverage at $n = 100$ (Section 2.6), and the sample size of MISTIE II is $n = 96$.

2.6 Simulation Studies

2.6.1 Simulations without a Baseline Variable

Two outcomes are separately considered: RICV5 and a binary outcome. No baseline variable is used. For RICV5, the data generating distributions under treatment and control are the empirical distributions in MISTIE. No restrictions are made. The bounds are $(\psi_l, \psi_u) = (0.82, 0.96)$.

For the binary outcome, the data generating distribution is $P_0(Y_T = y) = P_0(Y_C =$

$y) = 0.5$ for $y \in \{1, 2\}$. We consider the cases of no restrictions and the no harm assumption; the bounds are $(\psi_l, \psi_u) = (0, 0.5)$ and $(\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}) = (0, 0)$, respectively, where $\mathcal{R} = \{\text{no harm}\}$. We call these two cases binary (no restrictions) and binary (no harm).

For each case, we simulate 10,000 randomized trials each with n participants ($\frac{n}{2}$ in treatment, $\frac{n}{2}$ in control). We consider $n = 100, 500,$ and 1000, respectively. Using each simulated trial, the estimators from Section 2.4.2 are computed. Also, we compute a two-sided 95% CI for the lower bound and a separate two-sided 95% CI for the upper bound, using n-bootstrap and m-out-of-n bootstrap. For n-bootstrap, we generate 10,000 replicated data sets by resampling n participants, with replacement, from the simulated trial. The percentile method is used to get the 95% CI. For m-out-of-n bootstrap, we generate the 10,000 replicated data sets each by sampling m participants with replacement. The choice of m follows the procedure presented in Section 2.4.5.

Table 2.1 shows the empirical bias and standard error of the bound estimators for each case. Columns labeled “lower" give results for the lower bound estimator. Columns labeled “upper" give results for the upper bound estimator. Bias is negligible for RICV5. For the binary outcome, bias is substantial; the bias contribution to the mean squared error, as a percentage, ranges from 31 to 34%. The results for the lower bound in the no restrictions case and for both bounds in the no harm case are almost identical. This is because $\overline{\psi}_l$, $\overline{\psi}_l^{\mathcal{R}}$, and $\overline{\psi}_u^{\mathcal{R}}$ are identical if the outcome is binary and

Table 2.1: Bias and Standard Error of Proposed Estimator.

		Bias		Standard error	
Case	n	*lower*	*upper*	*lower*	*upper*
RICV5	100	0.002	0.000	0.055	0.027
	500	-0.000	-0.000	0.025	0.012
	1000	-0.000	-0.000	0.018	0.008
Binary	100	0.040	-0.040	0.059	0.059
(no restrictions)	500	0.018	-0.018	0.026	0.026
	1000	0.012	-0.013	0.018	0.018
Binary	100	0.040	0.040	0.058	0.058
(no harm)	500	0.018	0.018	0.026	0.026
	1000	0.013	0.013	0.018	0.018

$\mathcal{R} = \{\text{no harm}\}$, as proved below. Any differences in results between these estimators are due to sampling variability.

Proof *Consider a binary outcome whose two possible values are coded as 1 and 2, with 2 being better than 1. Let* $\mathcal{R} = \{$*no harm*$\}$*. In this proof, we will show that* $\overline{\psi}_l$*,* $\overline{\psi}_l^{\mathcal{R}}$*, and* $\overline{\psi}_u^{\mathcal{R}}$ *are equivalent.*

By Gadbury et al. (2004), we have $\overline{\psi}_l = \max\{0, \hat{p}_T(2) - \hat{p}_C(2)\}$*, where*

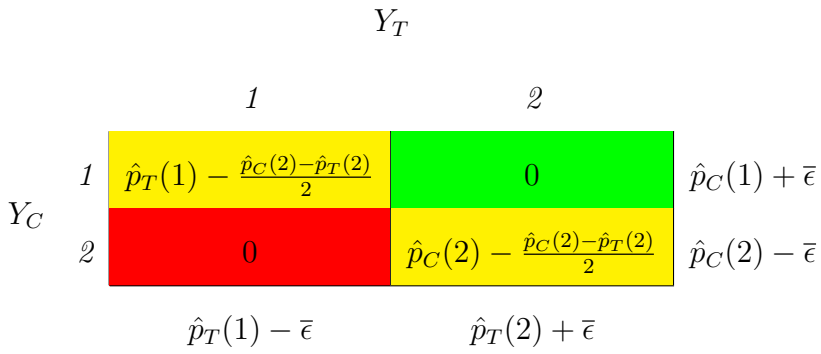
$$\begin{aligned} \widehat{p}_C(i) &= \frac{\sum_{m=1}^{n} 1(Y_m = i, A_m = 0)}{\sum_{m=1}^{n} 1(A_m = 0)}, \\ \widehat{p}_T(j) &= \frac{\sum_{m=1}^{n} 1(Y_m = j, A_m = 1)}{\sum_{m=1}^{n} 1(A_m = 1)}. \end{aligned}$$

We will show that $\overline{\psi}_l^{\mathcal{R}} = \overline{\psi}_u^{\mathcal{R}} = \max\{0, \hat{p}_T(2) - \hat{p}_C(2)\}$.

Case 1 *If* $\hat{p}_T(2) \geq \hat{p}_C(2)$*, it readily follows that* $\overline{\epsilon} = 0$ *and* $\overline{\psi}_l^{\mathcal{R}} = \overline{\psi}_u^{\mathcal{R}} = \hat{p}_T(2) - \hat{p}_C(2)$*. Since* $\hat{p}_T(2) \geq \hat{p}_C(2)$*, we have that* $\max\{0, \hat{p}_T(2) - \hat{p}_C(2)\} = \hat{p}_T(2) - \hat{p}_C(2)$*. Hence,*

$\overline{\psi}_l^{\mathcal{R}} = \overline{\psi}_u^{\mathcal{R}} = \max\{0, \hat{p}_T(2) - \hat{p}_C(2)\}$.

Case 2 *If $\hat{p}_T(2) < \hat{p}_C(2)$, it follows that $\overline{\epsilon} > 0$ because there exist no joint distributions that have marginal distributions $\hat{p}_T$ and $\hat{p}_C$ and satisfy the no harm assumption. It can be shown that $\overline{\epsilon} = \frac{\hat{p}_C(2) - \hat{p}_T(2)}{2}$. There exists one joint distribution which satisfies $|\pi_{1,1} + \pi_{1,2} - \hat{p}_C(1)| \leq \overline{\epsilon}$, $|\pi_{1,1} + \pi_{2,1} - \hat{p}_T(1)| \leq \overline{\epsilon}$, and the no harm assumption:*



It follows that $\overline{\psi}_l^{\mathcal{R}} = \overline{\psi}_u^{\mathcal{R}} = 0$. Since $\hat{p}_T(2) < \hat{p}_C(2)$, $\max\{0, \hat{p}_T(2) - \hat{p}_C(2)\} = 0$. Hence, $\overline{\psi}_l^{\mathcal{R}} = \overline{\psi}_u^{\mathcal{R}} = \max\{0, \hat{p}_T(2) - \hat{p}_C(2)\}$.

□

We compare the plug-in estimator to our estimator in the binary (no harm) case, in which they can differ due to the restriction. The plug-in estimator is undefined in 46% of simulations for $n = 100$, 48% for $n = 500$, and 49% for $n = 1000$. Conditional on being well-defined, it has bias 0.074 ($n = 100$), 0.034 ($n = 500$), 0.024 ($n = 1000$). Our estimator is less biased (Table 2.1) since it is equivalent to the plug-in estimator if the latter is well-defined, and is 0 (i.e., equal to the true lower and upper bounds) otherwise. Conditional on being well-defined, the plug-in estimator has standard

Table 2.2: Coverage Probabilities and Average Widths of CI's Constructed using n-bootstrap and m-out-of-n bootstrap.

		Coverage				Average width			
		n-bootstrap		m-out-of-n bootstrap		n-bootstrap		m-out-of-n bootstrap	
Case	n	*lower*	*upper*	*lower*	*upper*	*lower*	*upper*	*lower*	*upper*
RICV5	100	.937	.839	.943	.848	.205	.088	.218	.093
	500	.943	.941	.971	.947	.098	.046	.116	.049
	1000	.950	.948	.973	.960	.070	.033	.084	.037
Binary	100	.973	.898	.987	.930	.196	.243	.231	.278
(no restrictions)	500	.975	.893	.987	.934	.088	.109	.106	.129
	1000	.978	.886	.989	.927	.062	.077	.075	.092
Binary	100	.974	.974	.985	.985	.196	.196	.231	.231
(no harm)	500	.975	.975	.987	.987	.088	.088	.107	.107
	1000	0.974	.974	.988	.988	.062	.062	.076	.076

error 0.061 ($n = 100$), 0.027 ($n = 500$), 0.019 ($n = 1000$). Our estimator has similar standard errors (Table 2.1).

Table 2.2 shows the empirical coverage probability of the nominal 95% CI's constructed using m-out-of-n and n-bootstrap. The columns labeled "lower" give results for the CI's for the lower bound, and columns labeled "upper" give the results for the CI's for the upper bound. For the binary outcome, the empirical coverage is above 95% except for the upper bound in the no restrictions case, where coverage is as low as 92.7% for m-out-of-n bootstrap and 88.6% for n-bootstrap. For RICV5, empirical coverage is close to the nominal coverage, except the coverage rates for the upper bound are $\approx 84\%$ when $n = 100$. In our simulations, m-out-of-n bootstrap has higher coverage probability and average CI width than n-bootstrap. Fan and Park (2010) report the coverage probabilities of n- and m-out-of-n bootstrap both have

approximately the nominal coverage in simulations for their problem.

2.6.2 Simulation Studies with a Baseline Variable

We ran another set of simulations with a baseline variable to evaluate how subdividing into more strata affects the properties of our bound estimator. We consider two outcomes RICV5 and 180-day mRS. No restrictions are imposed for either outcome. For RICV5, the baseline variable is baseline clot volume. Three cases are considered, number of strata $K = 2$, 4, and 8. For $K = 8$, the eight strata are constructed using the octiles of the MISTIE II baseline clot volumes as cutpoints. The probability of belonging to each stratum is set as $1/8$. The marginal distributions under treatment and control for each stratum are set as the empirical distributions in MISTIE. We get $K = 4$ strata by merging every two adjacent strata of $K = 8$. We get $K = 2$ strata by merging every four adjacent strata of $K = 8$. The stratum probabilities and marginal distributions for $K = 2, 4$ are derived from those for $K = 8$. The bound parameters (ψ_l^X, ψ_u^X) are then (0.85, 0.96) for $K = 2$, (0.89, 0.96) for $K = 4$, and (0.89, 0.95) for $K = 8$. The bounds slightly narrow with higher K; the bound widths are 0.11 ($K = 2$), 0.07 ($K = 4$), and 0.06 ($K = 8$).

For 180-day mRS, the baseline variable is NIHSS score, which can be any integer from 1 to 42. We consider $K = 2$, 3, and 4. For $K = 4$, the four strata are minor stroke = 1-4, moderate stroke = 5-15, moderate/severe stroke = 16-20, and severe stroke = 21-42 (Kreutzer et al., 2011). The stratum probabilities are set to be those

in MISTIE, i.e., 0.03, 0.12, 0.27, and 0.57, respectively. For each of these $K = 4$ strata, the marginal distributions under treatment and control are set to be those observed in MISTIE. For $K = 3$, the three strata are obtained by merging the first two strata of $K = 4$. For $K = 2$, we merge the first three strata of $K = 4$. The stratum probabilities and marginal distributions for $K = 2, 3$ are derived using those for $K = 4$. The bound parameters (ψ_l^X, ψ_u^X) are then (0.14, 0.65) for $K = 2$, (0.16, 0.59) for $K = 3$, and (0.20, 0.59) for $K = 4$. The bounds narrow with higher K; the bound widths are 0.51, 0.43, and 0.39, respectively.

The simulations are implemented as follows. For a given outcome and value K, we simulate 10,000 randomized trials each with n participants $\left(\frac{n}{2}\right.$ in treatment, $\left.\frac{n}{2} \text{ in control}\right)$. We consider sample size $n = 100$, 500, and 1000, respectively. We generate a participant's stratum membership X by taking a random draw from the K strata, in concordance with the stratum probabilities. For a participant assigned to control (treatment), the observed outcome Y is then generated by taking a random draw from the marginal distribution under control (treatment) for the stratum to which the participant belongs. For each simulated trial, the estimators $\overline{\psi}_l^X$ and $\overline{\psi}_u^X$ from Section 2.4.2 are computed.

The results are shown in Table 2.3 (RICV5) and Table 2.4 (180-day mRS). Each table is organized in the following way. In the "Number of Strata" column, the number of strata K is noted in addition to the corresponding bound parameters (ψ_l^X, ψ_u^X). The "Percent with undefined estimates" column gives the percent of simulations with

Table 2.3: RICV5: Estimator Properties as a Function of Number of Strata.

Number of Strata	n	Percent with undefined estimates	Bias		Standard error	
			lower	*upper*	*lower*	*upper*
2	100	0	0.013	-0.000	0.047	0.027
(0.85, 0.96)	500	0	0.001	0.000	0.022	0.012
	1000	0	0.000	0.000	0.016	0.008
4	100	0	0.004	-0.001	0.043	0.027
(0.89, 0.96)	500	0	-0.000	-0.000	0.020	0.012
	1000	0	-0.000	0.000	0.014	0.008
8	100	1.9	0.011	-0.005	0.044	0.034
(0.89, 0.95)	500	0	0.002	0.000	0.019	0.014
	1000	0	0.001	-0.000	0.014	0.010

Table 2.4: 180-day mRS Score: Estimator Properties as a Function of Number of Strata.

Number of Strata	n	Percent with undefined estimates	Bias		Standard error	
			lower	*upper*	*lower*	*upper*
2	100	0	0.057	-0.007	0.073	0.070
(0.14,0.65)	500	0	0.021	0.000	0.035	0.033
	1000	0	0.013	0.000	0.025	0.024
3	100	0.05	0.052	-0.005	0.073	0.071
(0.16, 0.59)	500	0	0.013	-0.001	0.036	0.032
	1000	0	0.007	-0.000	0.026	0.022
4	100	32.9	0.040	-0.013	0.073	0.071
(0.20, 0.59)	500	0.07	0.014	-0.000	0.035	0.031
	1000	0	0.007	-0.001	0.025	0.022

undefined bound estimates. Bound estimates are undefined if no treatment or no control subjects are observed for some stratum. Bias and standard error are computed excluding the simulations with undefined bound estimates. Within the "Bias" and "Standard Error" columns, the "lower" and "upper" subcolumns specify whether the results are for the lower or upper bound estimator.

The percent of simulations with undefined estimates is small in most cases. However, over 30% of simulations had undefined estimates for the $K = 4$, $n = 100$ case of 180-day mRS. This is not surprising since the probability of being in the first of the $K = 4$ strata is only 0.03. With the small trial size of $n = 100$, it is likely to observe no treatment or no control subjects in the first stratum. Interestingly, the bias and standard error of our estimator (conditional on it being well-defined) are not adversely affected (and sometimes can be even better), when the baseline variable is discretized finely compared to coarsely. For RICV5, there is little bias for $K = 2$, 4, and 8. The standard errors are similar among the various K, except $K = 8$ has a slightly higher standard error at $n = 100$ than the other K. For 180-day mRS, the standard errors are also very similar among the various K. However, there are some small differences in bias. For the lower bound, there is non-negligible bias and $K = 2$ tends to have more bias than the higher values of K at fixed sample size. For the upper bound, there is very little bias for all K, but $K = 4$ has slightly more bias than the smaller values of K at $n = 100$. Bias, standard error, and the probability of undefined estimates may be highly dependent on the data generating distribution.

2.7 Discussion

In the MISTIE application, the interval corresponding to the lower and upper bound estimates is wide for the mRS outcome, and narrow for RICV. Depending on

the outcome, the proposed estimator of the bounds can be informative.

For 180-day mRS, we have $(\overline{\psi}_l, \overline{\psi}_u) = (0.10, 0.73)$ and $(\overline{\psi}_l^{\mathcal{R}}, \overline{\psi}_u^{\mathcal{R}}) = (0.10, 0.18)$ when $\mathcal{R} = \{\text{no harm}\}$. The latter bound estimates, though much closer together than in the former case due to the lower and upper bound estimates being much closer together, are only valid if the no harm assumption is true. This can be appropriate in some clinical settings but not in others. It is possible to generate evidence against the restriction being true by considering the value of $\bar{\epsilon}$. Though certain deviations from the restrictions may be detectable through $\bar{\epsilon}$, other deviations may not be. One may view the bound estimators under nested sets of more stringent restrictions as being a sensitivity analysis to examine how much information on the fraction who benefit would result under different types of assumptions on harm/benefit.

Our method can be applied to a continuous outcome that has been discretized. Discretization should be done such that a change from one level to the next is clinically meaningful. We focus on the case where there are relatively few levels compared to the sample size; it is an open problem to handle the case where the number of levels is not small relative to the sample size.

2.8 Software

Our code is available online at https://github.com/emhuang1/fraction-who-benefit. In the "demo" folder, we demonstrate how to analyze a simulated data set using the

code. The results from analyzing the simulated dataset are printed here.

```
> est1
[1] 0.1322825 0.7673190 0.0000000
> est2
[1] 0.1322825 0.2143657 0.0000000
> est3
[1] 0.1322825 0.7673190 0.0000000
> est4
[1] 0.2143657 0.2143657 0.0000000
> est5
[1] 0.12720555 0.73380326 0.17922078 0.73636364 0.00000000
0.07723013 0.73134328 0.00000000
> est6
[1] 0.125296456 0.210329557 0.175324675 0.262337662 0.001948052
 0.077230128 0.160360986 0.000000000
> est7
[1] 0.13021492 0.73380326 0.17922078 0.73636364 0.00000000
 0.08313086 0.73134328 0.00000000
> est8
[1] 0.202693194 0.210329557 0.246753247 0.262337662 0.001948052
0.160360986 0.160360986  0.000000000
> result
            lowerLimit upperLimit   m
lower bound 0.02649258  0.2451933 300
upper bound 0.60000000  0.8253968 120
```

The results "est1" through "est8" are the bound estimates for a simulated data set under eight different settings, described in the analyzeDataset.R code in the "demo" folder on GitHub. The printed output for "result" gives the 95% two-sided CI for ψ_l and the 95% two-sided CI for ψ_u, constructed using m-out-of-n bootstrap with the simulated data set.

Chapter 3

Constructing a Confidence Interval for the Fraction who Benefit Using a Randomized Trial

3.1 Abstract

The fraction who benefit from treatment is defined as the proportion of patients whose potential outcome under treatment is better than that under control. Statistical inference for this parameter is challenging since it is only partially identifiable, even in our context of a randomized trial. We propose and evaluate a new method for constructing a confidence interval for the fraction who benefit, when the outcome is ordinal-valued (with binary outcomes as a special case). This confidence interval

procedure is proved to be pointwise consistent. Our method does not require any assumptions about the joint distribution of the potential outcomes, although it has the flexibility to incorporate a wide range of user-defined assumptions. A potential advantage of our approach is that, unlike existing confidence interval methods for partially identified parameters (such as m-out-of-n bootstrap and subsampling), we do not need to select m or the subsample size, which is generally a challenging problem. Our method is based on a stochastic optimization technique involving a second order, asymptotic approximation that, to the best of our knowledge, has not been applied to biomedical studies. This approximation leads to statistics that are solutions to quadratic programs, and so they can be computed efficiently using existing optimization tools. In all of our simulations, our method attains the nominal coverage probability or higher, and can have substantially narrower average width compared to the m-out-of-n bootstrap. We also apply our method to a completed trial data set of a new surgical intervention for severe stroke.

3.2 Introduction

The fraction who benefit from treatment is the proportion of patients whose potential outcome under treatment is better than that under control. In other words, it is the proportion who would be better off with treatment. This fraction may be of interest to patients and care providers deciding between treatment and control. It

may also be informative to medical researchers; for example, a small fraction indicates that an exclusive subgroup benefits and resources should be devoted towards identifying it. We aim to draw inferences about the fraction who benefit, using a randomized trial.

In general, the fraction who benefit (sometimes abbreviated as "the fraction") is non-identifiable from observed data, even in the randomized trial context. This occurs because only one potential outcome can be observed per patient. Typically, identifiability of the fraction necessitates strong, untestable assumptions on the joint distribution of the potential outcomes, such as independence of the potential outcomes within a person. We do not require any assumptions on the joint distribution and only consider assumptions based on subject matter knowledge. Since the fraction is generally non-identifiable in this setting, constructing a confidence interval is a challenging problem.

An existing confidence interval procedure for our problem involves applying the m-out-of-n bootstrap to estimators of lower and upper bounds (which are identifiable) on the fraction who benefit. The m-out-of-n bootstrap is a generalization of the standard nonparametric bootstrap, where bootstrap replicate data sets are generated by resampling m patients with replacement for $m \leq n$. The m-out-of-n bootstrap is recommended because the bound estimators for our problem can be non-regular (Huang et al., 2017), and the standard bootstrap can be inconsistent in such cases. Another existing method for constructing confidence intervals is the subsampling

approach of Romano and Shaikh (2008). Subsampling is similar to the m-out-of-n bootstrap, except resampling is done without replacement. A challenge in using m-out-of-n bootstrap or subsampling is how to select m to achieve good performance. We propose a new confidence interval method that avoids having to select m.

Through simulation, we compare our method to the m-out-of-n bootstrap with respect to coverage probability and average width. In all cases, our method has coverage probability at or above the nominal level, while the m-out-of-n bootstrap sometimes has coverage probability below the nominal level. In some cases, our method achieves substantially narrower average width than the m-out-of-n bootstrap, e.g., reduction in average width of 40%. Our method has good coverage probability even in cases where the lower and upper bound parameters are non-differentiable functions of the marginal distributions under treatment and control, as shown in Section 3.6.

We apply our method to the CLEAR III (Clot Lysis: Evaluating Accelerated Resolution of Intraventricular Haemorrhage III) randomized trial of a new surgical treatment for stroke, which had a sample size of 500 patients (Hanley et al., 2017). Outcomes included disability measured by the modified Rankin Scale and death. As examples of the output of our confidence interval procedure, the 95% confidence interval for the fraction who benefit is [0.01,0.18] for the outcome 30-day mortality, [0.05,0.34] for 180-day mortality, [0,0.64] for 30-day disability, and [0.03,0.86] for 180-day disability.

CHAPTER 3. CONFIDENCE INTERVAL FOR THE FRACTION WHO BENEFIT

Our confidence interval procedure is based on representing the problem as a stochastic optimization problem. Stochastic optimization involves maximizing or minimizing the expected value of a function of unknown parameters and random variables, based on repeated observations of the random variables (data). As a simple example, M-estimators can be represented in terms of solving stochastic optimization problems (van der Vaart, 2000, Chapter 5). Our problem is substantially harder, since its formulation as a stochastic optimization problem involves a set of additional constraints on the parameter space (specifically, that the parameter lies within a polyhedron). When the optimal solution converges to a point on the boundary of the parameter space, the resulting statistics are generally not asymptotically normal; this rules out standard confidence interval procedures, many of which require asymptotic normality.

Shapiro et al. (2014) present general approaches for deriving the asymptotic distributions of such challenging stochastic optimization problems. To the best of our knowledge, these general approaches have not previously been used to solve problems arising in biomedical studies. We tailor one such approach to solve our problem, using a second order, asymptotic approximation of the objective function. We provide a self-contained proof of the validity of our method, which can be understood without requiring knowledge of stochastic optimization.

The statistic derived using the above approach can be computed using quadratic programming, i.e., minimizing a quadratic function of the data and parameters subject to linear equality and inequality constraints on the parameters. We used the

"quadprog" solver in MATLAB 2013B. The computing time of our method (after preprocessing) is independent of the sample size, but dependent on the width of the confidence interval and the number of levels for the ordinal outcome (larger width and fewer levels take less time). Each confidence interval in the CLEAR III application was computed between 4 and 8 minutes. This running time can be further reduced through parallelization.

Section 3.3 provides an overview of the previous work on the fraction. In Section 3.4, we describe the data generating distribution and state assumptions that are used throughout the paper. Our new method is presented in Section 3.5, including proofs of its asymptotic properties. We evaluate the method through simulation in Section 3.6. It is applied to the CLEAR III randomized trial in Section 3.7. Future work is discussed in Section 3.8.

3.3 Related Work

In general, the fraction who benefit is non-identifiable without making untestable assumptions about the joint distribution of the potential outcomes. There has been work on deriving and estimating bounds on this parameter in our context of ordinal outcomes (Borusyak, 2015; Lu et al., 2016; Huang et al., 2017).

To construct a confidence interval for the fraction who benefit, one could use the bound estimators proposed in Huang et al. (2017) and apply the m-out-of-n bootstrap

to them. Romano and Shaikh (2008) propose a general confidence interval method based on subsampling for partially identified parameters, such as the fraction who benefit. Under the subsampling condition (i) in Theorem 3.4 of their paper, Romano and Shaikh (2008) prove pointwise consistency of their confidence interval method. However, it is difficult to establish whether this condition holds in our problem.

Other parameters that contrast the distribution of an ordinal outcome under treatment versus control include the number needed to treat and the parameter in a responder analysis (Snapinn and Jiang, 2007). However, these parameters require that the ordinal outcome be dichotomized into "success" or "failure". For example, in one analysis of the CLEAR III trial, the modified Rankin Scale outcome was considered a "success" if it was in the range 0-3 (Hanley et al., 2017). The parameter of interest in a responder analysis is the difference between the population proportions who have a successful outcome under treatment versus control, where success can be a function of baseline variables. The number needed to treat is the reciprocal of this difference (Gordis, 2009). A downside to dichotomization of the outcome is that improvements not crossing the dichotomization threshold are ignored. The fraction who benefit considers the full ordinal scale.

3.4 Notation, Parameter Definition, and Assumptions

3.4.1 Parameter Definition

Consider an ordinal outcome with a finite number of levels, L. Without loss of generality, we assume that the levels are numbered as integers from 1 to L, in order of least to most favorable. Denote Y_T as the potential outcome under treatment and Y_C as the potential outcome under control. Let P_0 denote the true, unknown joint distribution on (Y_C, Y_T). Let $\pi_{i,j}$ denote the probability that $Y_C = i$ and $Y_T = j$, i.e., $\pi_{i,j} = P_0(Y_C = i, Y_T = j)$. We say that a patient benefits from treatment compared to control if her/his potential outcome pair (y_C, y_T) satisfies $y_T > y_C$. She/he is harmed if $y_T < y_C$ and experiences no individual treatment effect if $y_T = y_C$. The fraction who benefit from treatment, our parameter of interest, is:

$$\psi_0 = P_0(Y_T > Y_C) = \sum_{j>i} \pi_{i,j}. \tag{3.1}$$

We propose a method to construct a confidence interval for the parameter ψ_0, which does not require assumptions about the joint distribution P_0. The method can incorporate restrictions on the support of P_0, supplied by the user based on subject matter knowledge. Support restrictions are assumptions that certain potential outcome pairs

(i, j) are not possible, i.e., $\pi_{i,j} = 0$. The no harm assumption ($\pi_{i,j} = 0$ if $i > j$) is one example. For conciseness, we refer to support restrictions as restrictions. The user specifies restrictions through a function $g : \mathcal{L} \times \mathcal{L} \to \{0, 1\}$, where $\mathcal{L}$ is the set of integers from 1 to L. For any given input (i, j), the user sets $g(i, j)$ to 0 if she/he assumes that $\pi_{i,j} = 0$, and 1 otherwise. If no restrictions are made, the function g outputs 1 for all inputs. Let $\mathcal{R}$ be the set of all joint distributions P on (Y_C, Y_T) that satisfy the restrictions:

$$\mathcal{R} = \{P \text{ on } (Y_C, Y_T) : P(Y_C = i, Y_T = j) = 0 \text{ if } g(i, j) = 0\}. \tag{3.2}$$

Assumption 3 *The user-defined support restrictions are correct, i.e., $P_0 \in \mathcal{R}$.*

Incorrect assumptions can lead to poor coverage probability of our method and the m-out-of-n bootstrap, as shown in Section 3.6.

3.4.2 Observed Data

We construct our confidence interval using data from a randomized trial. Let n be the number of participants in the trial. For each participant m, let A_m and Y_m denote the participant's treatment assignment (1 if treatment and 0 if control) and observed outcome, respectively. We assume that the vectors (A_m, Y_m), $i = m, .., n$, are fully observed. Other assumptions include the following:

Assumption 4 *For each participant m, her/his potential outcome pair $(Y_{C,m}, Y_{T,m})$ is an independent, identically distributed draw from the unknown joint distribution P_0.*

Assumption 5 *The treatment assignments, A_m, $m = 1, .., n$, are independent, identically distributed Bernoulli(θ), where $0 < \theta < 1$. The treatment assignments $\{A_m\}_{m=1}^n$ are independent of the potential outcome pairs $\{(Y_{C,m}, Y_{T,m})\}_{m=1}^n$.*

Assumption 6 *For each participant m, we have $Y_m = A_m Y_{T,m} + (1 - A_m) Y_{C,m}$.*

Assumption 5 is satisfied by a simple randomized trial design (Friedman et al., 2010). The value θ is the probability of being assigned to treatment, which is known and should not be 0 or 1. Assumption 6 connects observed outcomes to potential outcomes and is called the consistency assumption.

3.4.3 Non-identifiability of the Fraction who Benefit

The assumptions above imply that the vectors (A_m, Y_m), $m = 1, .., n$, are independent and identically distributed. Let (A, Y) denote the random vector corresponding to a generic participant in the randomized trial. The vector (A, Y) for each participant is called the observed data, to distinguish it from the vector of potential outcomes (Y_C, Y_T) which is partially unobserved. Let P_{obs} denote the population distribution on the observed data vector (A, Y). By Assumption 5, we have $P_{obs}(A = a) = \theta^a (1 - \theta)^{1-a}$. Let the vector $\boldsymbol{\gamma}^* = (\gamma_{01}^*, .., \gamma_{0L}^*, \gamma_{11}^*, .., \gamma_{1L}^*)$ denote the

marginal distributions of the potential outcomes under treatment and under control, where $\gamma^*_{0y} = P_0(Y_C = y)$ and $\gamma^*_{1y} = P_0(Y_T = y)$ for all y in $\mathcal{L}$. By Assumptions 5 and 6, we have that for all $y \in \mathcal{L}$:

$$\begin{aligned} \gamma^*_{0y} = P_0(Y_C = y) &= P_{obs}(Y = y|A = 0), \\ \gamma^*_{1y} = P_0(Y_T = y) &= P_{obs}(Y = y|A = 1). \end{aligned} \tag{3.3}$$

This implies that the marginal distributions of the potential outcomes are identifiable.

Because only one potential outcome is observed per participant, the fraction who benefit ψ_0 is typically non-identifiable from observed data. However, the marginal distributions $\boldsymbol{\gamma}^*$ and restrictions $\mathcal{R}$ may rule out certain possibilities. Let $\psi^{\mathcal{R}}_l(P_{obs})$ and $\psi^{\mathcal{R}}_u(P_{obs})$ denote the sharp lower and upper bounds on the fraction, given the marginal distributions and restrictions, i.e.,

$$\psi^{\mathcal{R}}_l(P_{obs}) = \min\{P(Y_T > Y_C) : P \text{ has marginal distributions equal to } \boldsymbol{\gamma}^* \text{ and } P \in \mathcal{R}\},$$

$$\psi^{\mathcal{R}}_u(P_{obs}) = \max\{P(Y_T > Y_C) : P \text{ has marginal distributions equal to } \boldsymbol{\gamma}^* \text{ and } P \in \mathcal{R}\}.$$

These bounds are functions of P_{obs} due to their dependency on $\boldsymbol{\gamma}^*$, and are identifiable because $\boldsymbol{\gamma}^*$ is identifiable. For conciseness, we suppress the dependency on P_{obs}. The bounds are discussed in Huang et al. (2017). The fraction who benefit ψ_0 must be between the bounds, i.e., $\psi_0 \in [\psi^{\mathcal{R}}_l, \psi^{\mathcal{R}}_u]$. Moreover, for any $\psi \in [\psi^{\mathcal{R}}_l, \psi^{\mathcal{R}}_u]$, there exists some joint distribution $P \in \mathcal{R}$ that has marginals $\boldsymbol{\gamma}^*$ and with fraction who benefit

$P(Y_T > Y_C)$ equal to ψ. Intuitively, the marginal distributions and restrictions rule out candidates outside of the range $[\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}]$ but candidates inside the range are not ruled out.

We use the following definition for pointwise consistency from Romano and Shaikh (2008) tailored to our problem.

Definition 1 *A confidence set CS_n for ψ_0 is pointwise consistent at level $1-\alpha$ if, for any data generating distribution P_{obs} on (A, Y), we have for all $\psi \in \left[\psi_l^{\mathcal{R}}(P_{obs}), \psi_u^{\mathcal{R}}(P_{obs})\right]$:*

$$\liminf_{n\to\infty} P_{obs}(\psi \in CS_n) \geq 1 - \alpha. \tag{3.4}$$

Pointwise consistency is that, if one were to consider an arbitrary data generating distribution P_{obs} on (A, Y), then for all $\psi \in [\psi_l^{\mathcal{R}}(P_{obs}), \psi_u^{\mathcal{R}}(P_{obs})]$, the confidence set CS_n includes ψ with at least $1 - \alpha$ probability when n is large. This is a desired property because the fraction who benefit ψ_0 must be somewhere in the range $[\psi_l^{\mathcal{R}}(P_{obs}), \psi_u^{\mathcal{R}}(P_{obs})]$ and the observed data distribution provides no information on where it lies within that range.

3.5 Proposed Method

We construct a 95% confidence set for the fraction who benefit ψ_0 through hypothesis test inversion. We consider candidate values of ψ on a grid on $[0, 1]$. In our simulations and data application (Sections 3.6 and 3.7), the grid that is used has

a point at every hundredth, i.e., $\psi = 0, 0.01, 0.02, ..., 1$. A candidate value of ψ is excluded from the confidence set if and only if the hypothesis test for ψ rejects. If the confidence set is not an interval, we form a confidence interval using the smallest and largest points of the set. We present our hypothesis test in Section 3.5.1 and provide its implementation in Section 3.5.2. The MATLAB code is also provided with this paper. The asymptotic properties of the resulting confidence interval are presented in Section 3.5.3.

3.5.1 Hypothesis Test for Candidate Value of ψ

Let Π denote the set of all L by L matrices with nonnegative, real-valued entries that sum to 1. Define the set $\Gamma \in \mathbb{R}^{2L}$ as

$$\Gamma = \left\{ \boldsymbol{\gamma} = (\gamma_{01}, \ldots, \gamma_{0L}, \gamma_{11}, \ldots, \gamma_{1L})^t : \begin{array}{l} \text{For some } \boldsymbol{\pi} \in \Pi, \text{ we have} \\ \pi_{i,j} = 0 \text{ if } g(i,j) = 0 \\ \gamma_{0i} = \sum_{j=1}^{L} \pi_{i,j} \text{ for all } i \in \mathcal{L} \\ \gamma_{1j} = \sum_{i=1}^{L} \pi_{i,j} \text{ for all } j \in \mathcal{L} \end{array} \right\}. \quad (3.5)$$

This set is comprised of the pairs of marginal distributions (under treatment and under control) that are compatible with the restrictions. For example, if there are no restrictions, then Γ is the set of all vectors with nonnegative entries such that the sum of the first L entries equals 1 and the sum of the last L entries equals 1. If the no harm assumption is made and $L = 2$, the set Γ comprises all vectors with

nonnegative entries that satisfy $\gamma_{12} \geq \gamma_{02}$ and the constraint in the previous sentence. Under Assumption 3, the pair of true marginal distributions $\boldsymbol{\gamma}^*$ is in the set Γ.

Consider any candidate value of $\psi \in [0, 1]$. Define the set Γ^ψ as (3.17), but adding the constraint that $\sum_{j>i} \pi_{i,j} = \psi$ on the right hand side. The set Γ^ψ is comprised of the pairs of marginal distributions (under treatment and under control) that are compatible with both the restrictions $\mathcal{R}$ and the fraction who benefit being equal to ψ. Note that Γ and Γ^ψ are sets of vectors and not random. Each of these sets is a bounded, closed, convex polyhedron.

The null and alternative hypotheses for the candidate value of ψ are

$$H_0(\psi) \quad : \quad \boldsymbol{\gamma}^* \in \Gamma^\psi \tag{3.6}$$

$$H_a(\psi) \quad : \quad \boldsymbol{\gamma}^* \notin \Gamma^\psi. \tag{3.7}$$

The null hypothesis means that the pair of marginals $\boldsymbol{\gamma}^*$ (which is a function of P_{obs}) is compatible with both the restrictions $\mathcal{R}$ and fraction who benefit being equal to ψ. That is, there exists a joint distribution P on (Y_C, Y_T) such that its marginals equal $\boldsymbol{\gamma}^*$, it satisfies the restrictions, and the fraction who benefit $P(Y_T > Y_C)$ equals ψ. The null hypothesis is equivalent to $\psi \in [\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}]$, while the alternative hypothesis is equivalent to $\psi \notin [\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}]$. Intuitively, the null hypothesis means that the candidate value of ψ is not ruled out by the marginals $\boldsymbol{\gamma}^*$ and the restrictions $\mathcal{R}$.

Let $\boldsymbol{V} = (A, Y)$. We use the notation $P_{obs}X$ to denote the expectation of X with

respect to P_{obs}, the distribution on $\boldsymbol{V}$ induced by P_0 and Assumptions 5 and 6, as discussed in Section 3.4.3. Define

$$F(\boldsymbol{\gamma}, \boldsymbol{V}) = \sum_{a=0}^{1} \sum_{j=1}^{L} 1(A = a) \left\{1(Y = j) - \gamma_{aj}\right\}^2 .$$

The minimizer of $P_{obs}F(\boldsymbol{\gamma}, \boldsymbol{V})$ over $\boldsymbol{\gamma} \in \Gamma$ is unique and equal to γ^* defined in (3.3).

Define the test statistic as

$$T_{n,\psi} = n \left\{ \inf_{\boldsymbol{\gamma} \in \Gamma^{\psi}} P_n F(\boldsymbol{\gamma}, \boldsymbol{V}) - \inf_{\boldsymbol{\gamma} \in \Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V}) \right\}, \tag{3.8}$$

where P_n denotes the empirical distribution (that is, $P_n F(\boldsymbol{\gamma}, \boldsymbol{V}) = \frac{1}{n} \sum_{m=1}^{n} F(\boldsymbol{\gamma}, \boldsymbol{V}_m)$).

Let $\widehat{\boldsymbol{\gamma}}$ be the vector $(\widehat{\gamma}_{01}, \ldots, \widehat{\gamma}_{0L}, \widehat{\gamma}_{11}, \ldots, \widehat{\gamma}_{1L})$, where $\widehat{\gamma_{0i}} = P_n(A = 0, Y = i)/P_n(A = 0)$ and $\widehat{\gamma_{1j}} = P_n(A = 1, Y = j)/P_n(A = 1)$ for $i, j \in \mathcal{L}$. The vector $\widehat{\boldsymbol{\gamma}}$ represents the empirical marginal distributions of the potential outcomes under control and under treatment. Let Discrep be the following function of $\widehat{\boldsymbol{\gamma}}$ and a generic vector $\boldsymbol{\gamma}$ of length $2L$:

$$\text{Discrep}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) = \sum_{a=0}^{1} \sum_{j=1}^{L} \left[P_n 1(A = a)\right] \left(\gamma_{aj} - \widehat{\gamma}_{aj}\right)^2 .$$

The indicator function notation $1(S)$ equals 1 if S is true and 0 otherwise.

Lemma 3

$$T_{n,\psi} = n\left\{\inf_{\boldsymbol{\gamma}\in\Gamma^{\psi}} \mathit{Discrep}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}}) - \inf_{\boldsymbol{\gamma}\in\Gamma} \mathit{Discrep}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}})\right\}. \tag{3.9}$$

Lemma 3 is useful for interpreting the test statistic. The function Discrep$(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}})$ is a weighted sum of the squared differences between corresponding elements of the input $\boldsymbol{\gamma}$ and pair of empirical marginals $\widehat{\boldsymbol{\gamma}}$. Intuitively, Discrep$(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}})$ measures the discrepancy between $\boldsymbol{\gamma}$ and $\widehat{\boldsymbol{\gamma}}$, with higher values indicating more discrepancy. The test statistic $T_{n,\psi}$ compares the minimum discrepancy from the empirical marginals $\widehat{\boldsymbol{\gamma}}$ attained by $\boldsymbol{\gamma}\in\Gamma^{\psi}$ versus that attained by $\boldsymbol{\gamma}\in\Gamma$.

We will reject the null hypothesis that $\boldsymbol{\gamma}^*\in\Gamma^{\psi}$ for large values of $T_{n,\psi}$, as described below. This involves computing the asymptotic distribution of the statistic under the null hypothesis, and rejecting if $T_{n,\psi}$ exceeds the 0.95 quantile of this distribution.

Let $\boldsymbol{W} = (W_{01},\ldots,W_{0L},W_{11},\ldots,W_{1L})^t \in \mathbb{R}^{2L}$ be a random (column) vector with $W_{aj} = 2\times 1(A=a)\left\{1(Y=j)-\gamma^*_{aj}\right\}$. Let $\boldsymbol{Z} = (Z_{01},\ldots,Z_{0L},Z_{11},\ldots,Z_{1L})^t \in \mathbb{R}^{2L}$ be a random (column) vector having a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} = P_0\boldsymbol{W}\boldsymbol{W}^t$. Define $C(\boldsymbol{\gamma}^*)$ and $C^{\psi}(\boldsymbol{\gamma}^*)$ as

$$C(\boldsymbol{\gamma}^*) = \{r(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*) : \boldsymbol{\gamma}\in\Gamma, r\in\mathbb{R}_+\}, \qquad C^{\psi}(\boldsymbol{\gamma}^*) = \{r(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*) : \boldsymbol{\gamma}\in\Gamma^{\psi}, r\in\mathbb{R}_+\},$$

where $\mathbb{R}_+$ is the set of nonnegative real numbers.

Theorem 4 *Under the null hypothesis $\boldsymbol{\gamma}^* \in \Gamma^\psi$, $T_{n,\psi}$ converges in distribution to T_ψ defined as*

$$T_\psi = \min_{\boldsymbol{h} \in C^\psi(\boldsymbol{\gamma}^*)} (\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2) - \min_{\boldsymbol{h} \in C(\boldsymbol{\gamma}^*)} (\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2). \tag{3.10}$$

Proof *We use the general argument in Section 5.1.3 of Shapiro et al. (2014), except tailored to our specific problem. The proof here is self-contained. The null hypothesis $\gamma^* \in \Gamma^\psi$ implies that the minimizer γ^* of $\min_{\gamma \in \Gamma} P_0 F(\gamma, V)$ is unique and satisfies $\gamma^*_{aj} = P_0(Y = j | A = a)$ for each $a \in \{0,1\}, j \in \{1, \ldots, L\}$. This implies $\nabla P_0 F(\gamma^*, V) = 0$.*

Define $Z_n = n^{1/2}\{\nabla P_n F(\gamma^*, V) - \nabla P_0 F(\gamma^*, V)\}$, where the gradient is with respect to γ^*. It follows that $Z_n = (Z_{01,n}, \ldots, Z_{0L,n}, Z_{11,n}, \ldots, Z_{1L,n})^t$, where

$$Z_{aj,n} = -2n^{1/2} P_n \left[1(A = a) \left\{ 1(Y = j) - \gamma^*_{aj} \right\} \right],$$

for each $a \in \{0,1\}, j \in \{1, \ldots, L\}$. By the multivariate central limit theorem, Z_n converges in distribution to Z defined above. Let D_n denote the $2L \times 2L$ diagonal matrix with first L diagonal elements equal to $2P_n 1(A = 0)$ and last L diagonal elements equal to $2P_n 1(A = 1)$. For simplicity, we assume that $\theta = 1/2$, i.e., $P_{obs}(A = a) = 1/2$ for each $a \in \{0,1\}$. It follows that (Z_n, D_n) converges in distribution to (Z, D), for D the $2L \times 2L$ identity matrix.

We next show

$$n\left\{\inf_{\gamma\in\Gamma} P_nF(\gamma,V) - P_nF(\gamma^*,V)\right\} = \min_{h\in C_n(\gamma^*)}(h^tZ_n + h^tD_nh/2), \quad (3.11)$$

$$n\left\{\inf_{\gamma\in\Gamma^\psi} P_nF(\gamma,V) - P_nF(\gamma^*,V)\right\} = \min_{h\in C_n^\psi(\gamma^*)}(h^tZ_n + h^tD_nh/2), \quad (3.12)$$

for

$$C_n(\gamma^*) = \{n^{1/2}(\gamma-\gamma^*) : \gamma\in\Gamma\}, \qquad C_n^\psi(\gamma^*) = \{n^{1/2}(\gamma-\gamma^*) : \gamma\in\Gamma^\psi\}.$$

To show (3.11), we have

$$\begin{aligned}
& n\left\{\inf_{\gamma\in\Gamma} P_nF(\gamma,V) - P_nF(\gamma^*,V)\right\} \\
= \;& n\inf_{\gamma\in\Gamma} P_n\{F(\gamma,V) - F(\gamma^*,V)\} \\
= \;& n\inf_{\gamma\in\Gamma}\sum_{a=0}^{1}\sum_{j=1}^{L} P_n1(A=a)\left[\{1(Y=j)-\gamma_{aj}\}^2 - \{1(Y=j)-\gamma^*_{aj}\}^2\right] \\
= \;& n\inf_{\gamma\in\Gamma}\sum_{a=0}^{1}\sum_{j=1}^{L} P_n1(A=a)\left[-2\{1(Y=j)-\gamma^*_{aj}\}(\gamma_{aj}-\gamma^*_{aj}) + (\gamma_{aj}-\gamma^*_{aj})^2\right] \\
= \;& \inf_{\gamma\in\Gamma}\left[n^{1/2}\sum_{a=0}^{1}\sum_{j=1}^{L} Z_{aj,n}(\gamma_{aj}-\gamma^*_{aj}) + \sum_{a=0}^{1} P_n1(A=a)\sum_{j=1}^{L}\{n^{1/2}(\gamma_{aj}-\gamma^*_{aj})\}^2\right] \\
= \;& \inf_{\gamma\in\Gamma}\left[n^{1/2}(\gamma-\gamma^*)^tZ_n + \sum_{a=0}^{1} P_n1(A=a)\sum_{j=1}^{L}\{n^{1/2}(\gamma_{aj}-\gamma^*_{aj})\}^2\right] \\
= \;& \min_{h\in C_n(\gamma^*)} h^tZ_n + h^tD_nh/2, \qquad (3.13)
\end{aligned}$$

which proves (3.11). The proof of (3.12) is analogous, except replacing Γ *by* Γ^ψ *and*

$C_n(\gamma^*)$ *by* $C_n^{\psi}(\gamma^*)$.

Taking the difference between the left sides of (3.12) and (3.11), we have

$$T_n = \min_{h \in C_n^{\psi}(\gamma^*)} (h^t Z_n + h^t D_n h/2) - \min_{h \in C_n(\gamma^*)} (h^t Z_n + h^t D_n h/2).$$

Since $C_n(\gamma^*) \uparrow C(\gamma^*)$, $C_n^{\psi}(\gamma^*) \uparrow C^{\psi}(\gamma^*)$, *and* (Z_n, D_n) *converges in distribution to* (Z, D), *it follows from the continuous mapping theorem that* T_n *converges in distribution to* T. □

The proof shows the connection between $T_{n,\psi}$ and T_{ψ}, under the null hypothesis. We derive that

$$n \left\{ \inf_{\boldsymbol{\gamma} \in \Gamma^{\psi}} P_n F(\boldsymbol{\gamma}, \boldsymbol{V}) - P_n F(\boldsymbol{\gamma}^*, \boldsymbol{V}) \right\} = \min_{\boldsymbol{h} \in C_n^{\psi}(\boldsymbol{\gamma}^*)} (\boldsymbol{h}^t \boldsymbol{Z}_n + \boldsymbol{h}^t \boldsymbol{D}_n \boldsymbol{h}/2), \quad (3.14)$$

$$n \left\{ \inf_{\boldsymbol{\gamma} \in \Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V}) - P_n F(\boldsymbol{\gamma}^*, \boldsymbol{V}) \right\} = \min_{\boldsymbol{h} \in C_n(\boldsymbol{\gamma}^*)} (\boldsymbol{h}^t \boldsymbol{Z}_n + \boldsymbol{h}^t \boldsymbol{D}_n \boldsymbol{h}/2), \quad (3.15)$$

where

$$C_n(\boldsymbol{\gamma}^*) = \{n^{1/2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*) : \boldsymbol{\gamma} \in \Gamma\}, \qquad C_n^{\psi}(\gamma^*) = \{n^{1/2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*) : \boldsymbol{\gamma} \in \Gamma^{\psi}\}.$$

The test statistic $T_{n,\psi}$ is the difference between the left sides of (3.14) and (3.15). The limit distribution T_{ψ}, in (3.10), is the difference of the right sides, except with $\boldsymbol{D}_n$, $\boldsymbol{Z}_n$, C_n, and C_n^{ψ} replaced with their limits. Their limits are the identity matrix, $\boldsymbol{Z}$, C, and C^{ψ}, respectively. Under the alternative hypothesis, the test statistic $T_{n,\psi}$

does not converge to T_ψ, since $C_n^\psi(\boldsymbol{\gamma}^*)$ does not converge to $C^\psi(\boldsymbol{\gamma}^*)$ in this case.

Theorem 5 *Under the alternative hypothesis $\boldsymbol{\gamma}^* \notin \Gamma^\psi$, for any $M \in \mathbb{R}$, $P(T_{n,\psi} > M) \to 1$.*

Proof *Assume the alternative hypothesis $H_a(\psi) : \boldsymbol{\gamma}^* \notin \Gamma^\psi$ holds. Choose any $M \in \mathbb{R}$. By Lemma 3,*

$$T_{n,\psi} = n\left\{\inf_{\boldsymbol{\gamma}\in\Gamma^\psi} Discrep_{\widehat{\theta}}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}}) - \inf_{\boldsymbol{\gamma}\in\Gamma} Discrep_{\widehat{\theta}}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}})\right\},$$

where $\widehat{\theta} = P_n A$ and $Discrep_{\widehat{\theta}}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}}) = \sum_{a=0}^{1}\sum_{j=1}^{L}\left[(\gamma_{aj}-\widehat{\gamma}_{aj})^2\left(\widehat{\theta}\right)^a\left(1-\widehat{\theta}\right)^{1-a}\right]$.

Let $Discrep_\theta(\boldsymbol{\gamma},\boldsymbol{\gamma}^*) = \sum_{a=0}^{1}\sum_{j=1}^{L}\left[\left(\gamma_{aj}-\gamma_{aj}^*\right)^2\theta^a(1-\theta)^{1-a}\right]$. We have as $n\to\infty$

$$\left(\widehat{\boldsymbol{\gamma}},\widehat{\theta}\right) \xrightarrow{p} (\boldsymbol{\gamma}^*,\theta),$$

by the Weak Law of Large Numbers, Slutsky's lemma, and Theorem 2.7(vi) in van der Vaart (2000).

For any given $n \in \mathbb{N}$, we have

$$\begin{aligned} P(T_{n,\psi} > M) &= P\left(n\left\{\inf_{\boldsymbol{\gamma}\in\Gamma^\psi} Discrep_{\widehat{\theta}}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}}) - \inf_{\boldsymbol{\gamma}\in\Gamma} Discrep_{\widehat{\theta}}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}})\right\} > M\right) \\ &= P\left(\inf_{\boldsymbol{\gamma}\in\Gamma^\psi} Discrep_{\widehat{\theta}}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}}) - \inf_{\boldsymbol{\gamma}\in\Gamma} Discrep_{\widehat{\theta}}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}}) > \frac{M}{n}\right). \quad (3.16) \end{aligned}$$

For conciseness, let

$$d_{n,\psi} = \inf_{\boldsymbol{\gamma}\in\Gamma^{\psi}} Discrep_{\widehat{\theta}}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}}) - \inf_{\boldsymbol{\gamma}\in\Gamma} Discrep_{\widehat{\theta}}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}}).$$

Thus, we have $P(T_{n,\psi} > M) = P(d_{n,\psi} > M/n)$. *By Lemma 4, Lemma 5, and the Continuous Mapping Theorem, we have* $\inf_{\boldsymbol{\gamma}\in\Gamma^{\psi}} Discrep_{\widehat{\theta}}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}}) \xrightarrow{p} \inf_{\boldsymbol{\gamma}\in\Gamma^{\psi}} Discrep_{\theta}(\boldsymbol{\gamma},\boldsymbol{\gamma}^*)$ *and* $\inf_{\boldsymbol{\gamma}\in\Gamma} Discrep_{\widehat{\theta}}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}}) \xrightarrow{p} \inf_{\boldsymbol{\gamma}\in\Gamma} Discrep_{\theta}(\boldsymbol{\gamma},\boldsymbol{\gamma}^*)$. *Let* $c = \inf_{\boldsymbol{\gamma}\in\Gamma^{\psi}} Discrep_{\theta}(\boldsymbol{\gamma},\boldsymbol{\gamma}^*)$ *and* $b = \inf_{\boldsymbol{\gamma}\in\Gamma} Discrep_{\theta}(\boldsymbol{\gamma},\boldsymbol{\gamma}^*)$. *If* $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$, $Discrep_{\theta}(\boldsymbol{\gamma},\boldsymbol{\gamma}^*) = 0$ *because* $\gamma_{aj} = \gamma^*_{aj}$ *for all* (a,j) *pairs. If* $\boldsymbol{\gamma} \neq \boldsymbol{\gamma}^*$, *we have* $Discrep_{\theta}(\boldsymbol{\gamma},\boldsymbol{\gamma}^*) > 0$ *since* $\gamma_{aj} \neq \gamma^*_{aj}$ *for some* (a,j) *pair and* $0 < \theta < 1$. *Since* $\boldsymbol{\gamma}^* \in \Gamma$, *we have that* $b = 0$. *We have* $c > 0$ *since* $\boldsymbol{\gamma}^* \notin \Gamma^{\psi}$ *and* Γ^{ψ} *is compact, which is proved in Lemma 4. By Slutsky's lemma, the random variable* $d_{n,\psi}$ *converges in probability to positive number* c. *Let* $\epsilon = c/100$. *Since the sequence* M/n *converges to 0 as* $n \to \infty$, *we have for sufficiently large* n,

$$P(d_{n,\psi} \geq c - \epsilon) \leq P(d_{n,\psi} > M/n) \leq 1.$$

Because $d_{n,\psi}$ *converges in probability to a, the probability on the left converges to 1. By the Squeeze Theorem,* $P(d_{n,\psi} > M/n) = P(T_{n,\psi} > M)$ *converges to 1.* □

Lemma 4 *The sets* Γ *and* Γ^{ψ}, $\psi \in [0,1]$, *are compact.*

Proof *Let ψ be an arbitrary number in $[0,1]$. The sets Γ and Γ^{ψ} are defined as:*

$$\Gamma = \left\{ \gamma = (\gamma_{01}, \ldots, \gamma_{0L}, \gamma_{11}, \ldots, \gamma_{1L})^t : \begin{array}{l} \pi_{i,j} \geq 0 \text{ for all } i,j \in \mathcal{L} \\ \sum_{i=1}^{L} \sum_{j=1}^{L} \pi_{i,j} = 1 \\ \pi_{i,j} = 0 \text{ if } g(i,j) = 0 \\ \gamma_{0i} = \sum_{j=1}^{L} \pi_{i,j} \text{ for all } i \in \mathcal{L} \\ \gamma_{1j} = \sum_{i=1}^{L} \pi_{i,j} \text{ for all } j \in \mathcal{L} \end{array} \right\}, \quad (3.17)$$

$$\Gamma^{\psi} = \left\{ \gamma = (\gamma_{01}, \ldots, \gamma_{0L}, \gamma_{11}, \ldots, \gamma_{1L})^t : \begin{array}{l} \pi_{i,j} \geq 0 \text{ for all } i,j \in \mathcal{L} \\ \sum_{i=1}^{L} \sum_{j=1}^{L} \pi_{i,j} = 1 \\ \pi_{i,j} = 0 \text{ if } g(i,j) = 0 \\ \gamma_{0i} = \sum_{j=1}^{L} \pi_{i,j} \text{ for all } i \in \mathcal{L} \\ \gamma_{1j} = \sum_{i=1}^{L} \pi_{i,j} \text{ for all } j \in \mathcal{L} \\ \sum_{j>i} \pi_{i,j} = \psi \end{array} \right\}. \quad (3.18)$$

We prove that the set Γ^{ψ} is compact. The proof for Γ is analogous. Define

$$\Pi^{\psi} = \left\{ \pi = (\pi_{1,1}, .., \pi_{1,L}, \pi_{2,1}, .., \pi_{2,L}, ..., \pi_{L,1}, .., \pi_{L,L})^t : \begin{array}{l} \pi_{i,j} \geq 0 \text{ for all } i,j \in \mathcal{L} \\ \sum_{i=1}^{L} \sum_{j=1}^{L} \pi_{i,j} = 1 \\ \pi_{i,j} = 0 \text{ if } g(i,j) = 0 \\ \sum_{j>i} \pi_{i,j} = \psi \end{array} \right\} \quad (3.19)$$

Choose any vector $\pi \in \Pi^{\psi}$. *By the definition of* Π^{ψ}, *the components of* π *satisfy* $0 \leq \pi_{i,j} \leq 1$ *for all* i, j. *Thus, we have*

$$||\pi|| = \sqrt{\pi_{1,1}^2 + \cdots + \pi_{1,L}^2 + \pi_{2,1}^2 + \cdots + \pi_{2,L}^2 + \cdots + \pi_{L,1}^2 + \cdots + \pi_{L,L}^2} \leq L.$$

It follows that the set Π^{ψ} *is bounded. Also, the set* Π^{ψ} *is closed since it is a polyhedron.*

Since Π^{ψ} *is closed and bounded, it is compact. Define the mapping* $F : \Pi^{\psi} \to \mathbb{R}^{2L}$, *where*

$$\begin{aligned} F(\pi) &= (\gamma_{01}, \ldots, \gamma_{0L}, \gamma_{11}, \ldots, \gamma_{1L}), \\ \gamma_{0i} &= \sum_{j=1}^{L} \pi_{i,j} \text{ for all } i \in \mathcal{L}, \\ \gamma_{1j} &= \sum_{i=1}^{L} \pi_{i,j} \text{ for all } j \in \mathcal{L}. \end{aligned}$$

The mapping F *is continuous by Proposition 11.1, Theorem 11.2, and Theorem 11.4 in Fitzpatrick (1996). Let* $F(\Pi^{\psi})$ *denote the image of* $F : \Pi^{\psi} \to \mathbb{R}^{2L}$, *i.e.,*

$$F(\Pi^{\psi}) = \{\gamma | \gamma = F(\pi) \text{ for some point } \pi \in \Pi^{\psi}\}.$$

By Theorem 11.12 in Fitzpatrick (1996), $F(\Pi^{\psi})$ *is compact. Since* $\Gamma^{\psi} = F(\Pi^{\psi})$, *the set* Γ^{ψ} *is compact.* □

Lemma 5 *The functions* $g : (0,1) \times \mathbb{R}^{2L} \to \mathbb{R}$ *and* $g_\psi : (0,1) \times \mathbb{R}^{2L} \to \mathbb{R}$*, defined as*

$$g(\widetilde{\theta}, \widetilde{\gamma}) = \inf_{\gamma \in \Gamma} Discrep_{\widetilde{\theta}}(\gamma, \widetilde{\gamma}), \tag{3.20}$$

$$g_\psi(\widetilde{\theta}, \widetilde{\gamma}) = \inf_{\gamma \in \Gamma^\psi} Discrep_{\widetilde{\theta}}(\gamma, \widetilde{\gamma}), \tag{3.21}$$

are continuous at (θ, γ^*)*.*

Proof *We prove that* g_ψ *is continuous at* (θ, γ^*)*. The proof for* g *is analogous. Consider any sequence* $(\widetilde{\theta}_n, \widetilde{\gamma}_n)$ *that converges to* (θ, γ^*)*, where* $(\widetilde{\theta}_n, \widetilde{\gamma}_n) \in (0,1) \times \mathbb{R}^{2L}$ *for every* n*. We want to show that* $g_\psi(\widetilde{\theta}_n, \widetilde{\gamma}_n)$ *converges to* $g_\psi(\theta, \gamma^*)$*.*

Define the mappings $f : \mathbb{R}^{2L} \to \mathbb{R}$ *and* $f_n : \mathbb{R}^{2L} \to \mathbb{R}$ *as:*

$$f(\gamma) = Discrep_\theta(\gamma, \gamma^*),$$

$$f_n(\gamma) = Discrep_{\widetilde{\theta}_n}(\gamma, \widetilde{\gamma}_n).$$

Thus, we have

$$g_\psi(\theta, \gamma^*) = \inf_{\gamma \in \Gamma^\psi} f(\gamma),$$

$$g_\psi(\widetilde{\theta}_n, \widetilde{\gamma}_n) = \inf_{\gamma \in \Gamma^\psi} f_n(\gamma).$$

Consider any $\gamma \in \Gamma^{\psi}$. *Let* $\delta_n(\gamma) = f_n(\gamma) - f(\gamma)$. *Then we have*

$$\begin{aligned}\delta_n(\gamma) &= (1-\theta)\sum_{j=1}^{L}(\gamma_{0j}^* - \widetilde{\gamma}_{n,0j})^2 + \theta\sum_{j=1}^{L}(\gamma_{1j}^* - \widetilde{\gamma}_{n,1j})^2 \\ &\quad +(\theta - \widetilde{\theta}_n)\sum_{j=1}^{L}(\gamma_{0j} - \widetilde{\gamma}_{n,0j})^2 + (\widetilde{\theta}_n - \theta)\sum_{j=1}^{L}(\gamma_{1j} - \widetilde{\gamma}_{n,1j})^2 \\ &\quad +2(1-\theta)\sum_{j=1}^{L}(\gamma_{0j} - \gamma_{0j}^*)(\gamma_{0j}^* - \widetilde{\gamma}_{n,0j}) + 2\theta\sum_{j=1}^{L}(\gamma_{1j} - \gamma_{1j}^*)(\gamma_{1j}^* - \widetilde{\gamma}_{n,1j}).\end{aligned}$$

Let α *be an arbitrary positive number. Since* $(\widetilde{\theta}_n, \widetilde{\gamma}_n)$ *converges to* (θ, γ^*), *there exists a positive integer* N_α *such that if* $n \geq N_\alpha$

$$(\widetilde{\theta}_n - \theta)^2 + \sum_{j=1}^{L}(\gamma_{0j}^* - \widetilde{\gamma}_{n,0j})^2 + \sum_{j=1}^{L}(\gamma_{1j}^* - \widetilde{\gamma}_{n,1j})^2 < \alpha^2.$$

Thus, for $n \geq N_\alpha$,

$$\begin{aligned}
|\delta_n(\gamma)| &< (1-\theta)\alpha^2 + \theta\alpha^2 + |\theta - \widetilde{\theta}_n| \sum_{j=1}^{L} (\gamma_{0j} - \widetilde{\gamma}_{n,0j})^2 + |\widetilde{\theta}_n - \theta| \sum_{j=1}^{L} (\gamma_{1j} - \widetilde{\gamma}_{n,1j})^2 \\
&\quad + 2(1-\theta) \sum_{j=1}^{L} |\gamma_{0j} - \gamma^*_{0j}||\gamma^*_{0j} - \widetilde{\gamma}_{n,0j}| + 2\theta \sum_{j=1}^{L} |\gamma_{1j} - \gamma^*_{1j}||\gamma^*_{1j} - \widetilde{\gamma}_{n,1j}| \\
&< (1-\theta)\alpha^2 + \theta\alpha^2 + \alpha \sum_{j=1}^{L} (\gamma_{0j} - \widetilde{\gamma}_{n,0j})^2 + \alpha \sum_{j=1}^{L} (\gamma_{1j} - \widetilde{\gamma}_{n,1j})^2 \\
&\quad + 2(1-\theta)\alpha \sum_{j=1}^{L} |\gamma_{0j} - \gamma^*_{0j}| + 2\theta\alpha \sum_{j=1}^{L} |\gamma_{1j} - \gamma^*_{1j}| \\
&\leq (1-\theta)\alpha^2 + \theta\alpha^2 + \alpha \left[\sup_{\gamma \in \Gamma^\psi} \sum_{j=1}^{L} (\gamma_{0j} - \widetilde{\gamma}_{n,0j})^2 \right] + \alpha \left[\sup_{\gamma \in \Gamma^\psi} \sum_{j=1}^{L} (\gamma_{1j} - \widetilde{\gamma}_{n,1j})^2 \right] \\
&\quad + 2(1-\theta)\alpha \left[\sup_{\gamma \in \Gamma^\psi} \sum_{j=1}^{L} |\gamma_{0j} - \gamma^*_{0j}| \right] + 2\theta\alpha \left[\sup_{\gamma \in \Gamma^\psi} \sum_{j=1}^{L} |\gamma_{1j} - \gamma^*_{1j}| \right].
\end{aligned}$$

The set Γ^ψ *is compact by Lemma 4. Let* h_1, h_2, h_3, h_4 *be mappings from* $\Gamma^\psi \to \mathbb{R}$ *defined as:*

$$h_1(\gamma) = \sum_{j=1}^{L} (\gamma_{0j} - \widetilde{\gamma}_{n,0j})^2, \quad h_2(\gamma) = \sum_{j=1}^{L} (\gamma_{1j} - \widetilde{\gamma}_{n,1j})^2,$$
$$h_3(\gamma) = \sum_{j=1}^{L} |\gamma_{0j} - \gamma^*_{0j}|, \quad h_4(\gamma) = \sum_{j=1}^{L} |\gamma_{1j} - \gamma^*_{1j}|.$$

These mappings are continuous. For any given $j \in \{1,2,3,4\}$, *by the Extreme Value Theorem, the function* $h_j : \Gamma^\psi \to \mathbb{R}$ *attains a largest value, i.e.,* $\sup_{\gamma \in \Gamma^\psi} h_j(\gamma) = h_j(\overline{\gamma}_j)$ *for some* $\overline{\gamma}_j \in \Gamma^\psi$. *Let* $M_j = h_j(\overline{\gamma}_j)$. *Let* $M = \max\{M_j : j \in \{1,2,3,4\}\}$.

Thus, we have for all $n \geq N_\alpha$,

$$0 \leq |\delta_n(\gamma)| < (1-\theta)\alpha^2 + \theta\alpha^2 + \alpha M + \alpha M + 2(1-\theta)\alpha M + 2\theta\alpha M = \alpha^2 + 4\alpha M.$$

Since this result holds for arbitrary $\gamma \in \Gamma^\psi$, *it follows that if* $n \geq N_\alpha$,

$$0 \leq \sup_{\gamma \in \Gamma^\psi} |\delta_n(\gamma)| < \alpha^2 + 4\alpha M.$$

Since the choice of α *was arbitrary, we have that* $\sup_{\gamma\in\Gamma^\psi} |\delta_n(\gamma)| \to 0$.

We want to show that $g_\psi(\widetilde{\theta}_n, \widetilde{\gamma}_n) \to g_\psi(\theta, \gamma^*)$. *For any given* n, *we have*

$$\begin{aligned} |g_\psi(\widetilde{\theta}_n, \widetilde{\gamma}_n) - g_\psi(\theta, \gamma^*)| &= \left| \inf_{\gamma\in\Gamma^\psi} f_n(\gamma) - \inf_{\gamma\in\Gamma^\psi} f(\gamma) \right| \\ &= \left| \inf_{\gamma\in\Gamma^\psi} \{f(\gamma) + f_n(\gamma) - f(\gamma)\} - \inf_{\gamma\in\Gamma^\psi} f(\gamma) \right|. \end{aligned}$$

It can be shown that

$$\inf_{\gamma\in\Gamma^\psi} f(\gamma) + \inf_{\gamma\in\Gamma^\psi} \{f_n(\gamma) - f(\gamma)\} \leq \inf_{\gamma\in\Gamma^\psi} \{f(\gamma) + f_n(\gamma) - f(\gamma)\} \leq \inf_{\gamma\in\Gamma^\psi} f(\gamma) + \sup_{\gamma\in\Gamma^\psi} \{f_n(\gamma) - f(\gamma)\}.$$

Consequently, we have

$$\inf_{\gamma\in\Gamma^\psi} \{f_n(\gamma) - f(\gamma)\} \leq \inf_{\gamma\in\Gamma^\psi} \{f(\gamma) + f_n(\gamma) - f(\gamma)\} - \inf_{\gamma\in\Gamma^\psi} f(\gamma) \leq \sup_{\gamma\in\Gamma^\psi} \{f_n(\gamma) - f(\gamma)\}.$$

It follows that

$$\begin{aligned}\left|\inf_{\gamma\in\Gamma^\psi} f_n(\gamma) - \inf_{\gamma\in\Gamma^\psi} f(\gamma)\right| &\leq \max\left\{\left|\inf_{\gamma\in\Gamma^\psi}\{f_n(\gamma)-f(\gamma)\}\right|, \left|\sup_{\gamma\in\Gamma^\psi}\{f_n(\gamma)-f(\gamma)\}\right|\right\} \\ &= \max\left\{\left|\inf_{\gamma\in\Gamma^\psi}\delta_n(\gamma)\right|, \left|\sup_{\gamma\in\Gamma^\psi}\delta_n(\gamma)\right|\right\} \\ &\leq \sup_{\gamma\in\Gamma^\psi}|\delta_n(\gamma)|.\end{aligned}$$

The second inequality holds because $|\sup_{\gamma\in\Gamma^\psi}\delta_n(\gamma)| \leq \sup_{\gamma\in\Gamma^\psi}|\delta_n(\gamma)|$ *and because* $|\inf_{\gamma\in\Gamma^\psi}\delta_n(\gamma)| = |-\sup_{\gamma\in\Gamma^\psi}(-\delta_n(\gamma))| = |\sup_{\gamma\in\Gamma^\psi}(-\delta_n(\gamma))| \leq \sup_{\gamma\in\Gamma^\psi}|-\delta_n(\gamma)| = \sup_{\gamma\in\Gamma^\psi}|\delta_n(\gamma)|$.

So we have

$$0 \leq |g_\psi(\widetilde{\theta}_n, \widetilde{\gamma}_n) - g_\psi(\theta,\gamma^*)| \leq \sup_{\gamma\in\Gamma^\psi}|\delta_n(\gamma)|.$$

Since $\sup_{\gamma\in\Gamma^\psi}|\delta_n(\gamma)| \to 0$*, we have by the Squeeze Theorem that* $|g_\psi(\widetilde{\theta}_n, \widetilde{\gamma}_n) - g_\psi(\theta,\gamma^*)| \to 0$*. This implies that* $g_\psi(\widetilde{\theta}_n, \widetilde{\gamma}_n) \to g_\psi(\theta,\gamma^*)$. □

Intuitively, Theorem 5 is that the test statistic goes to infinity under the alternative hypothesis. Since the test statistic converges to a distribution (which we can simulate) under the null hypothesis but to infinity under the alternative hypothesis, our test can differentiate between the null and alternative hypotheses, as the sample size goes to infinity.

For any given ψ, let $t_\psi^{0.95}$ denote the 0.95 quantile of T_ψ. Reject the null hypothesis $\gamma^* \in \Gamma^\psi$ if and only if $T_{n,\psi} > t_\psi^{0.95} + \epsilon$, where $\epsilon = 10^{-10}$. The tiny perturbation ϵ is

required for the proof of pointwise consistency in Section 3.5.3. Let CS_n be the 95% confidence set constructed by inverting our hypothesis test, i.e.,

$$CS_n = \{\psi : T_{n,\psi} \leq t_\psi^{0.95} + \epsilon\}. \tag{3.22}$$

3.5.2 Using Quadratic Programming to Implement the Hypothesis Test

We present how to compute $T_{n,\psi}$ and estimate $t_\psi^{0.95}$. The test statistic $T_{n,\psi}$ can be computed from its form in (3.8) or (3.9). We present how to use (3.8). This requires solving two problems: $\inf_{\boldsymbol{\gamma}\in\Gamma^\psi} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$ and $\inf_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$. We show that each is the minimization of a quadratic function subject to a finite number of linear equality and inequality constraints. This is known as a quadratic program. Quadratic programs can be solved efficiently using existing softwares, such as MATLAB or CPLEX.

Consider the problem $\inf_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$. Define the following as the unknown variables: $\{\pi_{i,j} : i, j \in \mathcal{L}\}$, $\{\gamma_{0i} : i \in \mathcal{L}\}$, $\{\gamma_{1j} : j \in \mathcal{L}\}$. Let $\boldsymbol{H}_t$ denote the vector including all of these variables. The function to be minimized, $P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$, simplifies to

$$\sum_{a=0}^{1}\sum_{j=1}^{L}\left[P_n(A=a, Y=j) + \gamma_{aj}^2 P_n(A=a) - 2\gamma_{aj} P_n(A=a, Y=j)\right]. \tag{3.23}$$

Each term of the form $P_n(event)$ is a constant because it can be directly computed from the randomized trial. Note that (3.23) is a quadratic function of the variables $\boldsymbol{H}_t$. In the problem $\inf_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$, the function (3.23) is minimized under the constraint $\boldsymbol{\gamma} \in \Gamma$. By (3.17), $\boldsymbol{\gamma} \in \Gamma$ means that: $\pi_{i,j} \geq 0$ for all $i, j \in \mathcal{L}$, $\sum_{i,j} \pi_{i,j} = 1$, $\pi_{i,j} = 0$ if $g(i,j) = 0$, $\gamma_{0i} = \sum_{j=1}^{L} \pi_{i,j}$ for all $i \in \mathcal{L}$, $\gamma_{1j} = \sum_{i=1}^{L} \pi_{i,j}$ for all $j \in \mathcal{L}$. These are linear equality and inequality constraints on the variables $\boldsymbol{H}_t$. Therefore, one can solve $\inf_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$ by finding the minimum of (3.23) under the above constraints, using quadratic programming.

The other problem required to compute $T_{n,\psi}$ is $\inf_{\boldsymbol{\gamma}\in\Gamma^\psi} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$. Its corresponding quadratic program is the same as that for $\inf_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$, except with the additional linear constraint that $\sum_{j>i} \pi_{i,j} = \psi$.

We use simulation to estimate T_ψ, which is the limiting distribution of $T_{n,\psi}$ under the null hypothesis. Each draw from T_ψ is computed as follows. Let $\widehat{\boldsymbol{\gamma}}_{\mathcal{R}}$ denote the minimizer over $\boldsymbol{\gamma} \in \Gamma$ of $P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$, previously solved to get the test statistic $T_{n,\psi}$. The minimizer may not be unique. We simply let $\widehat{\boldsymbol{\gamma}}_{\mathcal{R}}$ be the minimizer that is returned by "quadprog" in MATLAB. Generate a random draw of $\boldsymbol{Z}$. This requires first estimating $\boldsymbol{\Sigma}$ by replacing $\boldsymbol{\gamma}^*$ by $\widehat{\boldsymbol{\gamma}}_{\mathcal{R}}$ and P_0 by P_n in the definition of $\boldsymbol{\Sigma}$. Next, solve the two quadratic programs in (3.10). To solve the second quadratic program $\min_{\boldsymbol{h}\in C(\boldsymbol{\gamma}^*)}(\boldsymbol{h}^t\boldsymbol{Z} + \boldsymbol{h}^t\boldsymbol{h}/2)$, define the following variables: $\{\pi_{ij} : i, j \in \mathcal{L}\}$, $\boldsymbol{h} = (h_{01}, \ldots, h_{0L}, h_{11}, \ldots, h_{1L})^t$, $\boldsymbol{\gamma} = (\gamma_{01}, \ldots, \gamma_{0L}, \gamma_{11}, \ldots, \gamma_{1L})^t$. Let $\boldsymbol{H}$ denote the vector including all of these variables. Define the linear constraints: $\pi_{ij} \geq 0$, $\pi_{i,j} = 0$ if

$g(i,j) = 0$, $\gamma_{0j} = \sum_{i=1}^{L} \pi_{ij}$, $\gamma_{1j} = \sum_{j=1}^{L} \pi_{ij}$, $\boldsymbol{h} = \boldsymbol{\gamma} - (\sum_{i,j} \pi_{ij})\widehat{\boldsymbol{\gamma}}_{\mathcal{R}}$ (note, this is a vector of equalities). Define the quadratic program to be $\min \boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2$, over the variables $\boldsymbol{H}$ and under the above linear constraints. To solve the first quadratic program $\min_{\boldsymbol{h} \in C^{\psi}(\boldsymbol{\gamma}^*)}(\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2)$, do as above but add the constraint: $\sum_{i<j} \pi_{ij} = \psi \sum_{i,j} \pi_{ij}$.

We take 1000 draws and compute their 0.95 quantile. Denote this quantile as $\widehat{t}_{\psi}^{0.95}$. The hat symbol is due to the finite number of draws and because we used an estimate of $\boldsymbol{\Sigma}$. Reject the null hypothesis $\gamma^* \in \Gamma^{\psi}$ if $T_{n,\psi} > \widehat{t}_{\psi}^{0.95} + \epsilon$, where $\epsilon = 10^{-10}$. The confidence set computed from this procedure is denoted as $\widehat{CS}_n$, i.e.,

$$\widehat{CS}_n = \{\psi : T_{n,\psi} \leq \widehat{t}_{\psi}^{0.95} + \epsilon\}. \tag{3.24}$$

As the sample size n goes to infinity, the estimate $\widehat{t}_{\psi}^{0.95}$ converges to $t_{\psi}^{0.95}$, and therefore $\widehat{CS}_n$ converges to CS_n.

3.5.3 Properties of Confidence Set and Corresponding Confidence Interval

Theorem 4 implies the following:

Theorem 6 *The confidence set CS_n is pointwise consistent at level 0.95.*

Proof *Consider an arbitrary data generating distribution P_{obs} on (A, Y). Suppose the underlying distribution P_0 on (Y_C, Y_T) satisfies Assumption 3. Choose any ψ that*

is consistent with the marginal distributions and restrictions, i.e., $\boldsymbol{\gamma}^* \in \Gamma^\psi$. *Then for all* $\epsilon > 0$*:*

$$\begin{aligned}
\liminf_{n\to\infty} P_{obs}(\psi \in CS_n) &= \liminf_{n\to\infty} P_{obs}(T_{n,\psi} \leq t_\psi^{0.95} + \epsilon) \\
&\geq \liminf_{n\to\infty} P_{obs}(T_{n,\psi} < t_\psi^{0.95} + \epsilon) \\
&\geq P_{obs}(T_\psi < t_\psi^{0.95} + \epsilon) \\
&\geq P_{obs}(T_\psi \leq t_\psi^{0.95}) \\
&= 0.95,
\end{aligned}$$

where the second inequality follows from Theorem 4 and the Portmanteau Lemma (van der Vaart, 2000). □

Theorem 5 implies the following:

Theorem 7 *For any* ψ *satisfying* $\boldsymbol{\gamma}^* \notin \Gamma^\psi$*, the probability that* ψ *is excluded from* CS_n *converges to 1.*

Proof *Consider an arbitrary data generating distribution* P_{obs} *on* (A, Y)*. Suppose the underlying distribution* P_0 *satisfies Assumption 3. Consider any* ψ *such that* $\boldsymbol{\gamma}^* \notin \Gamma^\psi$*. For any given* $\epsilon > 0$*,*

$$\lim_{n\to\infty} P_{obs}(\psi \notin CS_n) = \lim_{n\to\infty} P_{obs}(T_{n,\psi} > t_\psi^{0.95} + \epsilon) = 1.$$

□

The confidence set CS_n generated through test inversion is not guaranteed to be an interval. A confidence interval, denoted as CI_n, is constructed by taking the minimum and maximum of CS_n, i.e., $CI_n = [\min CS_n, \max CS_n]$. From pointwise consistency of the confidence set CS_n, the confidence interval CI_n is also pointwise consistent. We focus on the confidence interval because it is simpler to report, compared to the corresponding set.

Let $\widehat{CI}_n$ denote the confidence interval computed from $\widehat{CS}_n$, where

$$\widehat{CI}_n = [\min \widehat{CS}_n, \max \widehat{CS}_n].$$

To compute it efficiently, we implement the hypothesis test for $\psi = 0$ and for successively larger ψ only until failing to reject, in order to obtain the left endpoint of $\widehat{CI}_n$. To obtain the right endpoint, we implement the hypothesis test for $\psi = 1$ and for successively smaller ψ until failing to reject. This reduces computation time because the hypothesis test does not need to be done for every candidate value of ψ in the grid on $[0, 1]$. Wider intervals will take less time to run.

3.6 Simulation Studies

We use simulation to assess $\widehat{CI}_n$ at sample sizes n ranging from 200 to 2000. We compare it to the m-out-of-n bootstrap, with respect to coverage probability and average width. Let A and B denote the left and right endpoints of the confidence

interval constructed using the m-out-of-n bootstrap. To compute the value A, 10,000 bootstrap data sets are generated, each by sampling $m \leq n$ participants with replacement from the trial data set. Using each bootstrap data set, the lower and upper bounds $\psi_l^{\mathcal{R}}$ and $\psi_u^{\mathcal{R}}$ are estimated using the consistent estimators $\overline{\psi}_l^{\mathcal{R}}$ and $\overline{\psi}_u^{\mathcal{R}}$ proposed in Huang et al. (2017). These estimators were those defined in the previous chapter. For their intuition and the proof of consistency, refer to Huang et al. (2017). The value A is taken to be the 0.025 quantile of the 10,000 lower bound estimates. The value B is the 0.975 quantile of the 10,000 upper bound estimates. The rationale behind the choice of A and B is

$$\begin{aligned}
P_{obs}(A \leq \psi_0 \leq B) &\geq P_{obs}(A \leq \psi_l^{\mathcal{R}} \leq \psi_0 \leq \psi_u^{\mathcal{R}} \leq B) \\
&= P_{obs}(A \leq \psi_l^{\mathcal{R}} \leq \psi_u^{\mathcal{R}} \leq B) \\
&= 1 - P_{obs}(A > \psi_l^{\mathcal{R}} \text{ or } B < \psi_u^{\mathcal{R}}) \\
&\geq 1 - P_{obs}(A > \psi_l^{\mathcal{R}}) - P_{obs}(B < \psi_u^{\mathcal{R}}) \\
&\geq 1 - 0.025 - 0.025 \\
&= 0.95.
\end{aligned}$$

As a sensitivity analysis, we vary m between $m = n$, $0.9n$, $0.75n$, $0.5n$, and $0.25n$.

Table 3.1: Simulation Settings.

	L	γ^*_{01}	γ^*_{02}	γ^*_{11}	γ^*_{12}	User-defined restrictions $\mathcal{R}$	$(\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}})$
A	2	0.5	0.5	0.5	0.5	no restrictions	(0,0.5)
B	2	0.5	0.5	0.5	0.5	no harm	(0,0)
C	2	0.5	0.5	0.25	0.75	no restrictions	(0.25, 0.5)
D	6	empirical marginals in MISTIE II				no restrictions	(0.82,0.96)

3.6.1 Setup

We compare our method to the m-out-of-n bootstrap in four settings. The settings, labeled A-D, are outlined in Table 3.1. Each setting is a unique choice of the number of levels L, the marginal distributions $\boldsymbol{\gamma}^*$, the restrictions $\mathcal{R}$, and the true bound parameters $(\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}})$. For each setting, we conduct a simulation study at each of the following sample sizes: n = 200, 500, 1000, and 2000. For Settings A-C, each simulation study includes 5000 simulations. For Setting D, each study includes 1000 simulations since the six-level ordinal outcome results in longer running times. The randomization probability θ is set to 0.5 in all simulations. The steps to run a single simulation are as follows:

1. We generate a data set consisting of the treatment assignments and observed outcomes of n participants, i.e., (A_m, Y_m) with $m = 1, ..., n$. Each participant is randomly assigned to treatment or control using the randomization probability $\theta = 0.5$. Her/his observed outcome is a random draw from either $\boldsymbol{\gamma}^*_{0\cdot}$ or $\boldsymbol{\gamma}^*_{1\cdot}$, depending on the assigned treatment.

2. A 95% CI for the fraction who benefit is computed using the method we pro-

posed in Section 3.5.

3. A 95% CI for the fraction who benefit is computed using the m-out-of-n bootstrap, as described at the beginning of Section 3.6.

In each simulation study, we plot the coverage probability of each method as a mapping from $[0, 1]$ to $\mathbb{R}$. For any given $\psi \in [0, 1]$, the coverage probability of ψ equals the proportion of the confidence intervals that contain ψ. In addition, we compute the average width for each method.

3.6.2 Results

We present the coverage probabilities at $n = 500$ for Settings A and B in Figures 3.1 and 3.2, respectively. In each figure, we shade the region from $\psi = \psi_l^{\mathcal{R}}$ to $\psi = \psi_u^{\mathcal{R}}$ in grey. In Setting B, we have $(\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}) = (0, 0)$, so the grey region is the thin line at $\psi = 0$. In general, under Assumption 3, the fraction who benefit ψ_0 must be in the grey region and could be anywhere in this region. For ψ in the grey region, the probability that the confidence interval contains ψ should be ≥ 0.95.

In Figures 3.1 and 3.2, our method has coverage probabilities ≥ 0.95 for all ψ in the grey region. Moreover, our method achieves this in all four settings and at all sample sizes $n = 200, 500, 1000, 2000$. In contrast, the m-out-of-n bootstrap can have coverage probability < 0.95 in the grey region. This occurs in Setting A for the choices $m = n$, $m = 0.9n$ and $m = 0.75n$. For instance, the coverage probability of

$\psi = 0.5$, which lies in the grey region, is 0.95 for our method but 0.89 for m-out-of-n bootstrap ($m = n$). In contrast, for Settings B-D, the choice $m = n$ yields at least the nominal coverage and the smallest average width among the choices of m. This reflects the challenge of choosing m.

In Setting B (Figure 3.2), the set of ψ for which the null hypothesis $\gamma^* \in \Gamma^\psi$ is true is the single point 0 (under the no harm assumption). An impressive result is that, using our method, the confidence interval is $[0, 0]$ in 50% of the simulations. In other words, our method gives the best possible confidence interval 50% of the time, up to the precision of 0.01. The first point in the grid that should be excluded is $\psi = 0.01$. Our method excludes $\psi = 0.01$ 53% percent of the time. On the other hand, the m-out-of-n bootstrap excludes $\psi = 0.01$ only 6% of the time at best (with $m = n$). Our methods's ability to exclude ψ outside of the grey region translates to large improvements in average width.

Our method can have substantially shorter average width than the competitor. We observe this in Settings B and C. In Setting B, the reduction in average width of our method (compared to the m-out-of-n bootstrap) ranges from 37-69% at $n = 200$, 40-70% at $n = 500$, 41-71% at $n = 1000$, and 43-71% at $n = 2000$. (The ranges are due to trying different options for the choice of m.) In Setting C, the reduction in average width of our method ranges from 6-33% at $n = 200$, 7-32% at $n = 500$, 6-28% at $n = 1000$, and 6-23% at $n = 2000$.

In Settings A and D, the m-out-of-n bootstrap sometimes has narrower average

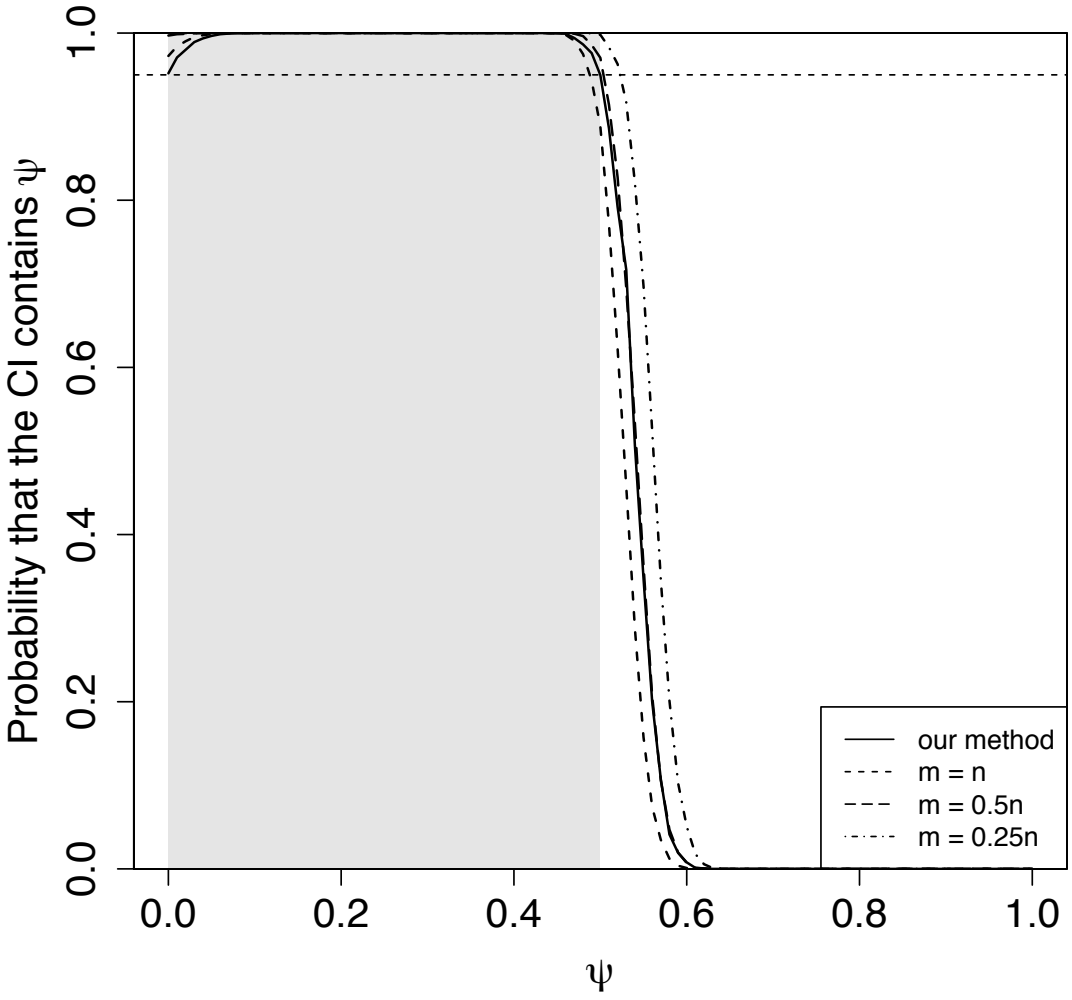


Figure 3.1: Coverage probabilities in Setting A at $n = 500$. *The grey region spans from $\psi = 0$ to $\psi = 0.5$, which are the lower and upper bounds $\psi_l^{\mathcal{R}}$ and $\psi_u^{\mathcal{R}}$ in Setting A. To achieve good coverage under Assumption 3, coverage probabilities should be ≥ 0.95 for all ψ in the grey region. For legibility of the plot, the curves for $m = 0.9n$ and $m = 0.75n$ are not shown. They lie between the curves for $m = n$ and $m = 0.5n$, but closely resemble the curve for $m = n$.*

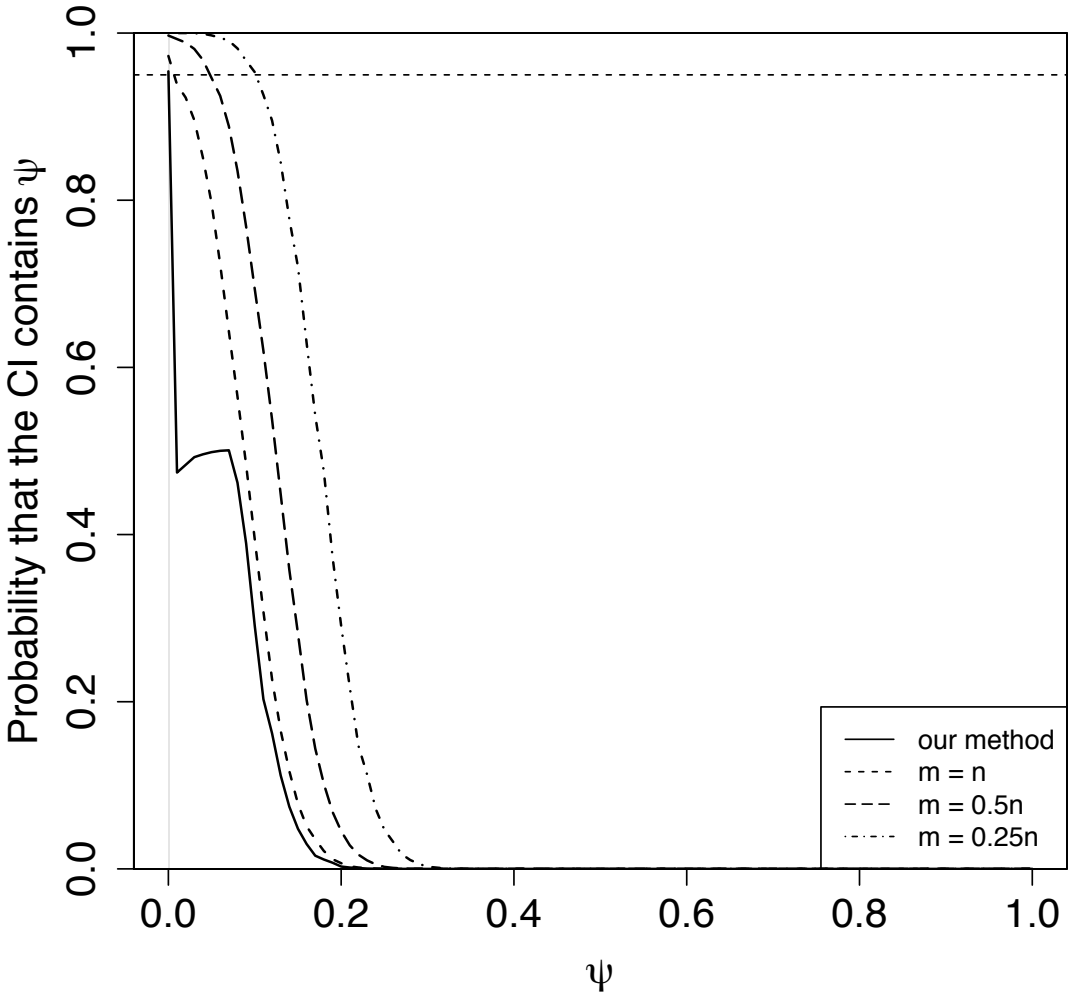


Figure 3.2: Coverage probabilities in Setting B at $n = 500$. *The grey region is the single point $\psi = 0$, since in Setting B the lower and upper bounds $\psi_l^{\mathcal{R}}$ and $\psi_u^{\mathcal{R}}$ are both zero. To achieve good coverage under Assumption 3, coverage probabilities should be ≥ 0.95 at $\psi = 0$. For legibility of the plot, the curves for $m = 0.9n$ and $m = 0.75n$ are not shown. They lie between the curves for $m = n$ and $m = 0.5n$, but closely resemble the curve for $m = n$.*

width than our method. In Setting A, this occurs only when the m-out-of-n bootstrap undercovers, i.e., has coverage probability < 0.95 in the grey region. In Setting D, the m-out-of-n bootstrap achieves narrower average width at $n = 200$, with an improvement ranging from 2-14%. At $n = 500$, the m-out-of-n bootstrap offers an improvement of 2% when $m = n$. However, our method has narrower average width at the higher sample sizes, with reductions in average width ranging from 3-23% at $n = 1000$ and 6-22% at $n = 2000$.

Our method and the m-out-of-n bootstrap can have poor coverage if Assumption 3 is violated. Consider Setting B and suppose that the no harm assumption does not hold. Then the true fraction who benefit could be larger than zero, but both our method and the m-out-of-n bootstrap have coverage probabilities below 0.95 for candidate values of $\psi > 0$ (Figure 3.2). To avoid violating Assumption 3, restrictions should either be based on subject matter knowledge or no restrictions should be made.

3.7 Application to the CLEAR III trial

3.7.1 Analysis Procedure

We apply our method to the CLEAR III (Clot Lysis: Evaluating Accelerated Resolution of Intraventricular Haemorrhage III) randomized trial (Hanley et al., 2017). This was a Phase III trial from 2009-2016 about intraventricular haemorrhage (IVH), which is bleeding into the ventricles of the brain, due to a stroke. CLEAR III tested

whether using the drug alteplase (treatment) to remove the blood clot from the ventricle results in a better functional outcome than using saline (control). The trial included 500 participants, with 249 assigned to alteplase and 251 to saline. The primary outcome was the modified Rankin scale (mRS) score at 180 days post-stroke. The mRS score is an ordinal rating of functional outcome with seven levels. The levels are defined as follows (this list is quoted directly from Cheng et al. (2014)):

0. no symptoms at all
1. no significant disability: despite symptoms, able to perform all usual duties and activities
2. slight disability: unable to perform all previous activities but able to look after own affairs without assistance
3. moderate disability: requiring some help but able to walk without assistance
4. moderately severe disability: unable to walk without assistance and unable to attend to own bodily needs without assistance
5. severe disability: bedridden, incontinent, and requiring constant nursing care and attention
6. death.

Based on CLEAR III, the proportion of patients with 180-day mRS ≤ 3 was estimated as 0.48 under alteplase and 0.45 under saline (95% CI for difference in proportions:

[-0.04, 0.12]).

We consider the primary outcome 180-day mRS, as well as the outcomes 30-day mRS, 30-day mortality, and 180-day mortality. For the mRS outcomes, we utilize the full ordinal scale. A separate analysis is performed for each outcome. First, a 95% CI for the fraction of patients who benefit from alteplase (relative to saline) is computed using the method proposed in Section 3.5. Participants whose outcome is missing are excluded. Out of 500 participants, the number of participants who are excluded is 6 for 30-day mRS, 9 for 180-day mRS, 0 for 30-day mortality, and 5 for 180-day mortality. Second, we attempt to answer the question of who benefits from alteplase compared to saline. An existing hypothesis was that patients with large baseline IVH volumes are more likely to benefit than patients with small baseline IVH volumes. We define 17.5 mL as the threshold between "small" and "large", per the suggestion of neurologist Daniel Hanley. By this definition, 188 of the 500 participants have small baseline volumes and 312 participants have large baseline volumes. We compute a 95% CI for the fraction of patients with a small baseline IVH volume who benefit from alteplase compared to saline. This is done by applying our method using solely the participants with baseline IVH clot volume $\leq$ 17.5 mL. The participants whose outcome is missing are excluded. Analogously, we compute a separate 95% CI for patients with a large baseline IVH volume.

All of the CI's are constructed without using any restrictions ($g = 1$), so Assumption 3 is met. In CLEAR III, simple randomization with $\theta = 0.5$ was implemented

for the first 100 participants. After this point, a covariate adaptive method was used to achieve balance between the alteplase and saline arms on two pre-selected baseline variables. Our method is currently designed for trials using simple randomization. For the purpose of demonstrating our method, we assume that simple randomization was performed throughout CLEAR III with $\theta = 0.5$. As future work, we will extend our method to handle more randomization schemes.

3.7.2 Results

The 95% CI's are presented in Table 3.2. For every CI, we specify the outcome of interest and whether the CI is for all patients, only those with baseline clot volume $\leq$ 17.5 mL, or only those with baseline clot volume $>$ 17.5 mL. The corresponding sample sizes are presented in Table 3.3. For every pairing of outcome with patient group, we present the sample sizes in the format, total (saline, alteplase).

Table 3.2: 95% CI's for the Fraction who Benefit from Alteplase Compared to Saline.

	All patients	$\leq$ 17.5 mL only	$>$ 17.5 mL only
30-day mRS	[0.00,0.64]	[0.00,0.79]	[0.00,0.61]
180-day mRS	[0.03,0.86]	[0.00,0.92]	[0.01,0.83]
30-day mortality	[0.01,0.18]	[0.03,0.13]	[0.00,0.23]
180-day mortality	[0.05,0.34]	[0.02,0.23]	[0.04,0.44]

We discuss the results for 30-day mortality. As shown in Table 3.2, the 95% CI for this outcome is [0.01, 0.18] for all patients. In words, we are 95% confident that the fraction of patients who benefit with respect to 30-day mortality (i.e., the proportion who would be alive under alteplase but dead under saline, at 30 days) is between

Table 3.3: Sample Sizes for 95% CI's in Table 3.2.

	All patients	$\leq$ 17.5 mL only	$>$ 17.5 mL only
30-day mRS	494 (249,245)	186 (95,91)	308 (154,154)
180-day mRS	491 (245,246)	185 (93,92)	306 (152,154)
30-day mortality	500 (251,249)	188 (95, 93)	312 (156,156)
180-day mortality	495 (247,248)	187 (94,93)	308 (153,155)

0.01 and 0.18. Our result contributes new knowledge that was not provided by the difference in proportions ATE. The proportion dead at 30 days was estimated to be 0.06 higher under saline compared to alteplase (95% CI: [-0.00,0.11]). The difference in proportions is a lower bound on the fraction who benefit. Using the 95% CI for the difference in proportions, we could only infer that the fraction who benefit is 0 or above. As shown in Table 3.2, the 95% CI for the fraction who benefit is $[0.03, 0.13]$ for those with baseline clot volume $\leq$ 17.5 mL, and $[0, 0.23]$ for those with baseline clot volume $>$ 17.5 mL. The first result suggests that there is a small proportion who benefit among those with $\leq$ 17.5 mL. However, it is inconclusive whether those with $>$ 17.5 mL are more likely to benefit than those with $\leq$ 17.5 mL since the 95% CI for $>$ 17.5 mL is wider in both directions. Further work is required to identify the subgroup that benefits with respect to 30-day mortality. One could develop a scalar score using multiple baseline covariates, and stratify patients using the score rather than a single baseline covariate.

As shown in Table 3.2, the 95% CI's for the mortality outcomes are narrow, while those for the mRS outcomes are very wide. The 95% CI for 180-day mRS and all patients is [0.03, 0.86]. The finite sample size of $n = 491$ (Table 3.3) contributes to

the CI width. However, we believe the main factor behind the wide width is that the true lower and upper bound parameters span a wide range. In other words, the marginal distributions of the potential outcomes alone are not very informative about the fraction who benefit. Support restrictions can potentially reduce the width of the bounds. In this setting, current subject matter knowledge does not imply any support restrictions.

3.8 Discussion

We have developed a new method for constructing a 95% confidence interval for the fraction who benefit. It offers the user flexibility to define support restrictions based on subject matter knowledge or to make no assumptions at all on the joint distribution of the potential outcomes. Our confidence interval is proved to be pointwise consistent. In simulation tests, it had empirical coverage of at least 95% when the sample size was varied from 200 to 2000. The method is computationally efficient because it uses quadratic programming to compute the test statistic and the distribution of the test statistic under the null hypothesis. It also avoids having to choose m.

Our simulations and CLEAR III application show that the confidence interval constructed using the method can be narrow and informative. However, we also encountered cases in which the confidence intervals are wide, likely due to the true lower

and upper bound parameters being far apart. Our confidence interval is designed so that for any given value between the lower and upper bounds, the coverage of the value is at least 0.95. Consequently, if the bounds are far apart, the confidence intervals will tend to be wide even for extremely large sample sizes. Huang et al. (2017) found that incorporating a baseline variable can substantially narrow the bounds, without requiring assumptions. As future work, we will incorporate a baseline variable into our method. Also, we will provide guidance on how to select a baseline variable that will be effective in tightening the bounds.

Chapter 4

Discussion

I have developed a method to estimate lower and upper bounds on the fraction who benefit. I prove that the estimator is consistent. It is also computationally efficient because it uses linear programming. In addition, I have developed a method to construct a confidence interval for the fraction. This confidence interval is pointwise consistent and can be computed using quadratic programming.

As future work, I will extend these methods to handle the case of missing outcome data due to loss to follow-up, which is common in real trials. The methods currently are designed for the case of a simple randomized trial design, i.e., each subject's treatment assignment is a random draw from the same Bernoulli distribution. I will extend the methods to more complex randomization schemes. In addition, I will extend the methods to handle not only randomized trial data, but also observational study data. This can be useful in applications where randomized trial data is unavailable. An

extension to observational studies would require some assumptions about how to get consistent estimates for the marginal distributions of the potential outcomes.

I plan to apply the methods to more randomized trials to determine whether the results tend to be informative (i.e., the bound estimates and the CI for the fraction who benefit are narrow). A challenge is addressing the case where the results are uninformative due to the bound parameters themselves being wide. To address this challenge, further work should try to incorporate subject matter knowledge that cannot be expressed in the form of support restrictions. One option is to make assumptions about the correlation between the potential outcomes under control and treatment. Other than user-defined assumptions, I prove in Chapter 2 that a baseline variable can tighten the bound parameters. Currently, the CI method in Chapter 3 only uses treatment assignment and outcome data from a trial. I will extend it to also incorporate data about a baseline variable. Because many baseline variables can be collected in a single trial, I will address the question of selecting which baseline variable to incorporate. The aim is to choose the variable that results in the most narrowing of the bound parameters.

Chapter 5

Appendix

Table 5.1: Estimated Bounds on the Fraction who Benefit, with respect to 30-day mRS Score.

	(a) For the population					**(b) For each subpopulation**					
	$\overline{\psi}_l$	$\overline{\psi}_u$	$\overline{\epsilon}$	$\overline{\psi}_l^X$	$\overline{\psi}_u^X$	$\overline{\psi}_{l,1}^X$	$\overline{\psi}_{u,1}^X$	$\overline{\epsilon}_1$	$\overline{\psi}_{l,2}^X$	$\overline{\psi}_{u,2}^X$	$\overline{\epsilon}_2$
No restrictions	.07	.61	0	.17	.52	.04	.41	0	.26	.60	0
	$\overline{\psi}_l^{\mathcal{R}}$	$\overline{\psi}_u^{\mathcal{R}}$	$\overline{\epsilon}$	$\overline{\psi}_l^{\mathcal{R},X}$	$\overline{\psi}_u^{\mathcal{R},X}$	$\overline{\psi}_{l,1}^{\mathcal{R},X}$	$\overline{\psi}_{u,1}^{\mathcal{R},X}$	$\overline{\epsilon}_1$	$\overline{\psi}_{l,2}^{\mathcal{R},X}$	$\overline{\psi}_{u,2}^{\mathcal{R},X}$	$\overline{\epsilon}_2$
Benefit ≤ 5	.07	.61	0	.17	.52	.04	.41	0	.26	.60	0
≤ 4	.07	.61	0	.17	.52	.04	.41	0	.26	.60	0
≤ 3	.07	.61	0	.17	.52	.04	.41	0	.26	.60	0
≤ 2	.09	.61	0	.17	.52	.04	.41	0	.26	.60	0
≤ 1	.12	.61	0	.22	.52	.06	.41	0	.33	.60	0
Harm ≤ 5	.07	.61	0	.17	.52	.04	.41	0	.26	.60	0
≤ 4	.07	.61	0	.17	.52	.04	.41	0	.26	.60	0
≤ 3	.07	.61	0	.17	.52	.04	.41	0	.26	.60	0
≤ 2	.07	.59	0	.17	.52	.04	.41	0	.26	.60	0
≤ 1	.07	.50	0	.17	.45	.04	.26	0	.26	.60	0
No Harm	.03	.26	.02	.15	.30	0	.26	.06	.26	.33	0

Table 5.2: Estimated Bounds on the Fraction who Benefit, with respect to 180-day mRS Score.

	(a) For the population					(b) For each subpopulation					
	$\overline{\psi}_l$	$\overline{\psi}_u$	$\overline{\epsilon}$	$\overline{\psi}_l^X$	$\overline{\psi}_u^X$	$\overline{\psi}_{l,1}^X$	$\overline{\psi}_{u,1}^X$	$\overline{\epsilon}_1$	$\overline{\psi}_{l,2}^X$	$\overline{\psi}_{u,2}^X$	$\overline{\epsilon}_2$
No restrictions	.10	.73	0	.13	.64	.01	.59	0	.22	.68	0
	$\overline{\psi}_l^{\mathcal{R}}$	$\overline{\psi}_u^{\mathcal{R}}$	$\overline{\epsilon}$	$\overline{\psi}_l^{\mathcal{R},X}$	$\overline{\psi}_u^{\mathcal{R},X}$	$\overline{\psi}_{l,1}^{\mathcal{R},X}$	$\overline{\psi}_{u,1}^{\mathcal{R},X}$	$\overline{\epsilon}_1$	$\overline{\psi}_{l,2}^{\mathcal{R},X}$	$\overline{\psi}_{u,2}^{\mathcal{R},X}$	$\overline{\epsilon}_2$
Benefit ≤ 5	.10	.73	0	.13	.64	.01	.59	0	.22	.68	0
≤ 4	.10	.73	0	.13	.64	.01	.59	0	.22	.68	0
≤ 3	.10	.73	0	.14	.64	.01	.59	0	.25	.68	0
≤ 2	.12	.73	0	.18	.64	.01	.59	0	.32	.68	0
≤ 1	.18	.64	0	.25	.64	.01	.57	0	.44	.69	.02
Harm ≤ 5	.10	.73	0	.13	.64	.01	.59	0	.22	.68	0
≤ 4	.10	.73	0	.13	.64	.01	.59	0	.22	.68	0
≤ 3	.10	.64	0	.13	.63	.01	.57	0	.22	.68	0
≤ 2	.10	.62	0	.13	.63	.01	.57	0	.22	.68	0
≤ 1	.10	.42	0	.12	.51	.00	.35	.04	.22	.64	0
No Harm	.10	.18	0	.12	.54	.00	.52	.08	.22	.55	0

Table 5.3: Estimated Bounds on the Fraction who Benefit, with respect to RICV5.

	(a) For the population					(b) For each subpopulation					
	$\overline{\psi}_l$	$\overline{\psi}_u$	$\overline{\epsilon}$	$\overline{\psi}_l^X$	$\overline{\psi}_u^X$	$\overline{\psi}_{l,1}^X$	$\overline{\psi}_{u,1}^X$	$\overline{\epsilon}_1$	$\overline{\psi}_{l,2}^X$	$\overline{\psi}_{u,2}^X$	$\overline{\epsilon}_2$
No restrictions	.82	.96	0	.83	.96	.81	.96	0	.86	.96	0
	$\overline{\psi}_l^{\mathcal{R}}$	$\overline{\psi}_u^{\mathcal{R}}$	$\overline{\epsilon}$	$\overline{\psi}_l^{\mathcal{R},X}$	$\overline{\psi}_u^{\mathcal{R},X}$	$\overline{\psi}_{l,1}^{\mathcal{R},X}$	$\overline{\psi}_{u,1}^{\mathcal{R},X}$	$\overline{\epsilon}_1$	$\overline{\psi}_{l,2}^{\mathcal{R},X}$	$\overline{\psi}_{u,2}^{\mathcal{R},X}$	$\overline{\epsilon}_2$
Benefit ≤ 4	.87	.96	0	.83	.98	.81	.96	0	.86	1	.05
≤ 3	.56	1	.18	.58	1	.68	1	.12	.47	1	.23
≤ 2	.27	1	.30	.33	1	.49	1	.22	.17	1	.38
≤ 1	.18	1	.35	.15	1	.30	1	.31	.00	1	.43
Harm ≤ 4	.82	.96	0	.83	.96	.81	.96	0	.86	.96	0
≤ 3	.82	.96	0	.83	.96	.81	.96	0	.86	.96	0
≤ 2	.82	.96	0	.83	.96	.81	.96	0	.86	.96	0
≤ 1	.82	.96	0	.83	.96	.81	.96	0	.86	.96	0
No Harm	.82	.96	0	.83	.96	.81	.96	0	.86	.96	0

Table 5.4: Estimated Bounds on the Fraction who Benefit, with respect to RICV2.

	(a) For the population					(b) For each subpopulation					
	$\overline{\psi}_l$	$\overline{\psi}_u$	$\overline{\epsilon}$	$\overline{\psi}_l^X$	$\overline{\psi}_u^X$	$\overline{\psi}_{l,1}^X$	$\overline{\psi}_{u,1}^X$	$\overline{\epsilon}_1$	$\overline{\psi}_{l,2}^X$	$\overline{\psi}_{u,2}^X$	$\overline{\epsilon}_2$
No restrictions	.84	.96	0	.87	.96	.84	.96	0	.89	.96	0
	$\overline{\psi}_l^{\mathcal{R}}$	$\overline{\psi}_u^{\mathcal{R}}$	$\overline{\epsilon}$	$\overline{\psi}_l^{\mathcal{R},X}$	$\overline{\psi}_u^{\mathcal{R},X}$	$\overline{\psi}_{l,1}^{\mathcal{R},X}$	$\overline{\psi}_{u,1}^{\mathcal{R},X}$	$\overline{\epsilon}_1$	$\overline{\psi}_{l,2}^{\mathcal{R},X}$	$\overline{\psi}_{u,2}^{\mathcal{R},X}$	$\overline{\epsilon}_2$
Benefit ≤ 10	.87	.96	0	.85	.98	.84	.96	0	.86	1	.05
≤ 9	.75	1	.10	.76	1	.90	.99	.03	.62	1	.15
≤ 8	.63	1	.15	.65	1	.68	1	.12	.62	1	.15
≤ 7	.52	1	.21	.49	1	.57	1	.18	.42	1	.25
≤ 6	.33	1	.29	.35	1	.52	1	.20	.17	1	.38
≤ 5	.28	1	.32	.28	1	.39	1	.25	.17	1	.38
≤ 4	.25	1	.33	.24	1	.36	1	.27	.12	1	.40
≤ 3	.22	1	.34	.17	1	.30	1	.31	.04	1	.43
≤ 2	.14	1	.37	.11	1	.19	1	.35	.04	1	.43
≤ 1	.06	1	.40	.10	1	.15	1	.38	.06	1	.43
Harm ≤ 10	.84	.96	0	.87	.96	.84	.96	0	.89	.96	0
≤ 9	.84	.96	0	.87	.96	.84	.96	0	.89	.96	0
≤ 8	.84	.96	0	.87	.96	.84	.96	0	.89	.96	0
≤ 7	.84	.96	0	.87	.96	.84	.96	0	.89	.96	0
≤ 6	.84	.96	0	.87	.96	.84	.96	0	.89	.96	0
≤ 5	.84	.96	0	.87	.96	.84	.96	0	.89	.96	0
≤ 4	.84	.96	0	.87	.96	.84	.96	0	.89	.96	0
≤ 3	.84	.96	0	.87	.96	.84	.96	0	.89	.96	0
≤ 2	.84	.96	0	.87	.96	.84	.96	0	.89	.96	0
≤ 1	.84	.96	0	.87	.96	.84	.96	0	.89	.96	0
No Harm	.84	.96	0	.87	.96	.84	.96	0	.89	.96	0

Table 5.5: Estimated Bounds on the Fraction who Benefit, with respect to RICV1.

	(a) For the population					(b) For each subpopulation					
	$\overline{\psi}_l$	$\overline{\psi}_u$	$\overline{\epsilon}$	$\overline{\psi}_l^X$	$\overline{\psi}_u^X$	$\overline{\psi}_{l,1}^X$	$\overline{\psi}_{u,1}^X$	$\overline{\epsilon}_1$	$\overline{\psi}_{l,2}^X$	$\overline{\psi}_{u,2}^X$	$\overline{\epsilon}_2$
No restrictions	.82	.96	0	.83	.96	.81	.96	0	.86	.96	0
	$\overline{\psi}_l^{\mathcal{R}}$	$\overline{\psi}_u^{\mathcal{R}}$	$\overline{\epsilon}$	$\overline{\psi}_l^{\mathcal{R},X}$	$\overline{\psi}_u^{\mathcal{R},X}$	$\overline{\psi}_{l,1}^{\mathcal{R},X}$	$\overline{\psi}_{u,1}^{\mathcal{R},X}$	$\overline{\epsilon}_1$	$\overline{\psi}_{l,2}^{\mathcal{R},X}$	$\overline{\psi}_{u,2}^{\mathcal{R},X}$	$\overline{\epsilon}_2$
Benefit ≤ 2	.87	.96	0	.83	.98	.81	.96	0	.86	1	.05
≤ 1	.27	1	.30	.28	1	.38	1	.21	.17	1	.38
Harm ≤ 2	.82	.96	0	.83	.96	.81	.96	0	.86	.96	0
≤ 1	.82	.96	0	.83	.96	.81	.96	0	.86	.96	0
No Harm	.82	.96	0	.83	.96	.81	.96	0	.86	.96	0

Table 5.6: Estimated Bounds on the Fraction who Benefit, with respect to RICV2.

	(a) For the population					(b) For each subpopulation					
	$\overline{\psi}_l$	$\overline{\psi}_u$	$\overline{\epsilon}$	$\overline{\psi}_l^X$	$\overline{\psi}_u^X$	$\overline{\psi}_{l,1}^X$	$\overline{\psi}_{u,1}^X$	$\overline{\epsilon}_1$	$\overline{\psi}_{l,2}^X$	$\overline{\psi}_{u,2}^X$	$\overline{\epsilon}_2$
No restrictions	.65	.91	0	.64	.83	.42	.70	0	.86	.96	0
	$\overline{\psi}_l^{\mathcal{R}}$	$\overline{\psi}_u^{\mathcal{R}}$	$\overline{\epsilon}$	$\overline{\psi}_l^{\mathcal{R},X}$	$\overline{\psi}_u^{\mathcal{R},X}$	$\overline{\psi}_{l,1}^{\mathcal{R},X}$	$\overline{\psi}_{u,1}^{\mathcal{R},X}$	$\overline{\epsilon}_1$	$\overline{\psi}_{l,2}^{\mathcal{R},X}$	$\overline{\psi}_{u,2}^{\mathcal{R},X}$	$\overline{\epsilon}_2$
Benefit ≤ 1	.87	.91	0	.76	.85	.66	.70	0	.86	1	.05
Harm ≤ 1	.65	.91	0	.64	.83	.42	.70	0	.86	.96	0
No Harm	.65	.87	0	.64	.81	.42	.66	0	.86	.96	0

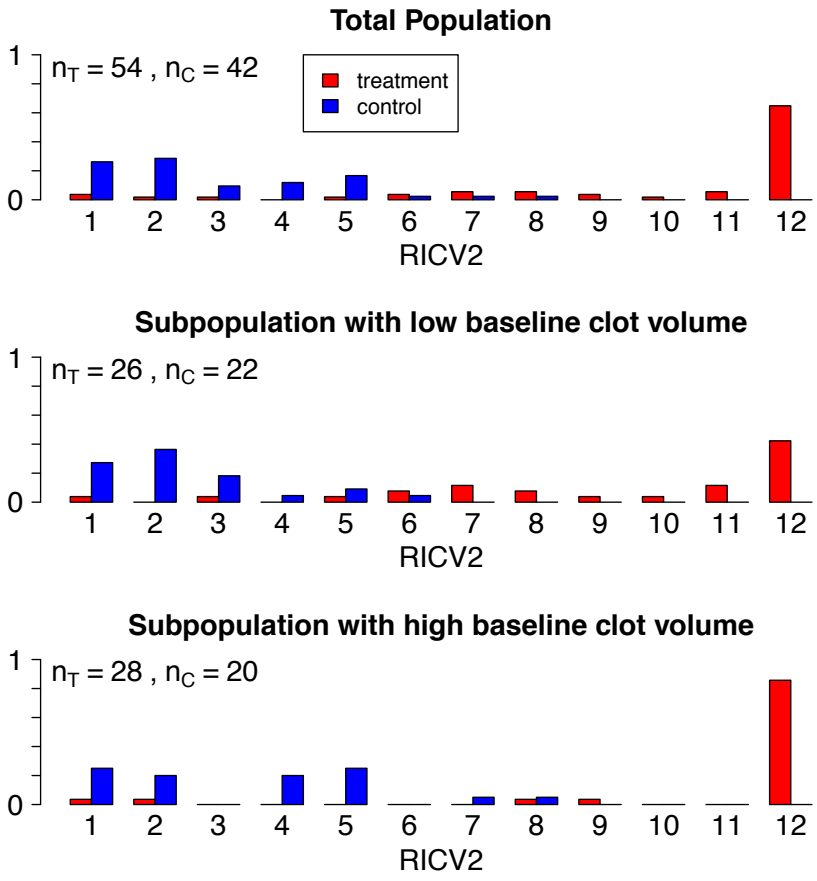


Figure 5.1: Empirical Probability Mass Functions of RICV2 under Treatment and Control.

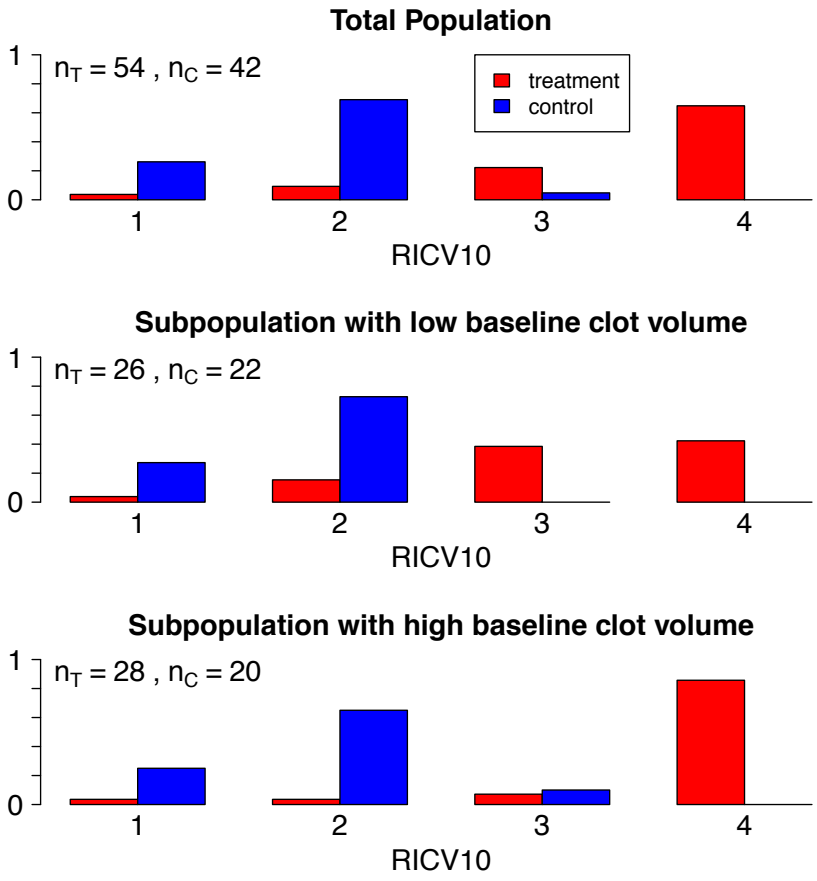


Figure 5.2: Empirical Probability Mass Functions of RICV10 under Treatment and Control.

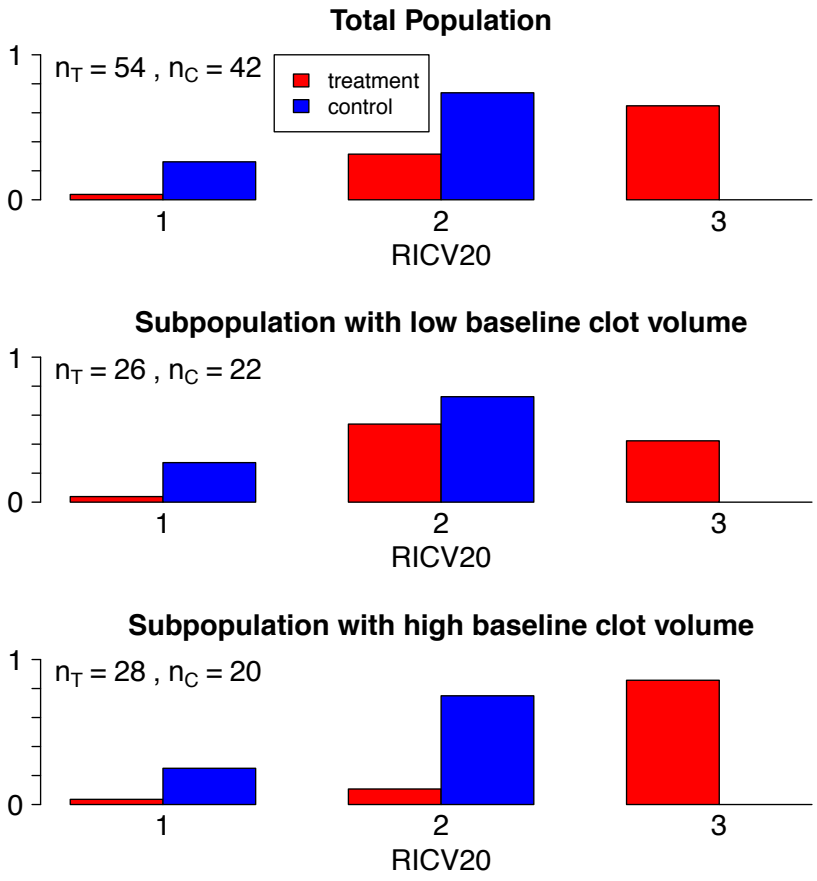


Figure 5.3: Empirical Probability Mass Functions of RICV20 under Treatment and Control.

Bibliography

Bickel, P., Götze, F., and van Zwet, W. (1997). Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statistica Sinica*, 7:1–31.

Bickel, P. and Sakov, A. (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, 18(3):967–985.

Borusyak, K. (2015). Bounding the population shares affected by treatments. *Technical Report: SSRN: http://ssrn.com/abstract=2473827*.

Cheng, B., Forkert, N., Zavaglia, M., Hilgetag, C., Golsari, A., Siemonsen, S., et al. (2014). Influence of stroke infarct location on functional outcome measured by the modified Rankin Scale. *Stroke*, 45(6):1695–1702.

Dantzig, G. B. and Thapa, M. (2006). *Linear Programming 2: Theory and Extensions*. Springer Science & Business Media.

Durrett, R. (2010). *Probability: Theory and Examples*. Cambridge University Press.

BIBLIOGRAPHY

Fan, Y. and Park, S. (2009). Partial identification of the distribution of treatment effects and its confidence sets. *Advances in Econometrics*, 25:3–70.

Fan, Y. and Park, S. (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 26(03):931–951.

Fitzpatrick, P. (1996). *Advanced Calculus: A Course in Mathematical Analysis*. PWS Publishing Company.

Friedman, L. M., Furberg, C. D., and DeMets, D. L. (2010). *Fundamentals of Clinical Trials Fourth Edition*. Springer.

Gadbury, G., Iyer, H., and Albert, J. (2004). Individual treatment effects in randomized trials with binary outcomes. *Journal of Statistical Planning and Inference*, 121(2):163–174.

Gordis, L. (2009). Epidemiology (fourth edition).

Hanley, D., Thompson, R., Muschelli, J., Rosenblum, M., McBee, N., Lane, K., Bistran-Hall, A., Mayo, S., Keyl, P., Gandhi, D., et al. (2016). Safety and efficacy of minimally invasive surgery plus alteplase in intracerebral haemorrhage evacuation (MISTIE): a randomised, controlled, open-label, phase 2 trial. *Lancet Neurol*, 15(12):1228–1237.

Hanley, D. F., Lane, K., McBee, N., Ziai, W., Tuhrim, S., Lees, K. R., Dawson, J., Gandhi, D., Ullman, N., Mould, A., et al. (2017). Thrombolytic removal of intra-

ventricular haemorrhage in treatment of severe stroke: results of the randomised, multicentre, multiregion, placebo-controlled clear iii trial. *Lancet*.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Huang, E. J., Fang, E. X., Hanley, D. F., and Rosenblum, M. (2017). Inequality in treatment benefits: Can we determine if a new treatment benefits the many or the few? *Biostatistics*, 18(2):308–324.

Kim, J. (2014). Identifying the distribution of treatment effects under support restrictions. *Technical Report: arXiv:1410.5885 [stat.ME]*.

Kreutzer, J., Caplan, B., and DeLuca, J. (2011). *Encyclopedia of Clinical Neuropsychology*. Springer New York.

Lu, J., Ding, P., and Dasgupta, T. (2016). Treatment effects on ordinal outcomes: Causal estimands and sharp bounds. *arXiv preprint arXiv:1507.01542*.

Manski, C. (1997). Monotone treatment response. *Econometrica*, 65(6):1311–1334.

Morgan, T., Zuccarello, M., Narayan, R., Keyl, P., Lane, K., and Hanley, D. (2008). Preliminary findings of the minimally-invasive surgery plus rtPA for intracerebral hemorrhage evacuation (MISTIE) clinical trial. *Acta neurochirurgica. Supplement*, 105:147–51.

BIBLIOGRAPHY

Mould, W., Carhuapoma, J., Muschelli, J., Lane, K., Morgan, T., McBee, N., et al. (2013). Minimally invasive surgery plus recombinant tissue-type plasminogen activator for intracerebral hemorrhage evacuation decreases perihematomal edema. *Stroke*, 44(3):627–634.

Quinn, T., Dawson, J., Walters, M., and Lees, K. (2009). Reliability of the modified Rankin Scale: A systematic review. *Stroke*, 40(10):3393–3395.

Rohde, C. (2014). *Introductory Statistical Inference with the Likelihood Function*. Springer.

Romano, J. P. and Shaikh, A. M. (2008). Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference*, 138(9):2786–2807.

Severini, T. (2005). *Elements of Distribution Theory*. Cambridge University Press.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2014). *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*. MOS-SIAM Series on Optimization.

Snapinn, S. M. and Jiang, Q. (2007). Responder analyses and the assessment of a clinically relevant treatment effect. *Trials*, 8:31–36.

van der Vaart, A. (2000). *Asymptotic statistics*, volume 3. Cambridge University Press.

BIBLIOGRAPHY

Vanderbei, R. (2013). *Linear Programming: Foundations and Extensions*. Springer.

Williamson, R. and Downs, T. (1990). Probabilistic arithmetic. I. Numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning*, 4(2):89–158.

Zhang, Z., Wang, C., Nie, L., and Soon, G. (2013). Assessing the heterogeneity of treatment effects via potential outcomes of individual patients. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(5):687–704.

Emily Huang

Johns Hopkins University
Department of Biostatistics
(410) 300-4660
ehuang19@jhu.edu
https://emhuangbiostat.wordpress.com/
https://github.com/emhuang1
Last updated on June 9, 2017

EDUCATION

PhD Biostatistics (Expected to graduate in Summer 2017) 2012-Present
Johns Hopkins University
Thesis Advisor: Dr. Michael Rosenblum
Dissertation Title: "Statistical Methods for the Fraction who Benefit using a Randomized Trial"

BSE Bioengineering and Mathematics 2008-2012
University of Pennsylvania
Magna Cum Laude. Academic advisors: Drs. Ravi Radhakrishnan and Philip Gressman

WORK EXPERIENCE

Predoctoral Trainee of Epidemiology and Biostatistics of Aging Training Program 2012-Present

Research Assistant at the Children's Hospital of Philadelphia 2009-2012
Supervisor: Dr. Christopher Bonafide

AWARDS

- 2016 ENAR Distinguished Student Paper Award
- 2015-Present Johns Hopkins Center of Excellence in Regulatory Science and Innovation Scholar
- 2014 Summer Institute in Statistics and Modeling in Infectious Diseases Tuition and Travel Scholarship
- 2008-2009, 2010-2011, 2011-2012 University of Pennsylvania Dean's List
- 2008-2011 Science Systems and Applications, Inc. Scholarship
- 2010 University of Pennsylvania School of Engineering Undergraduate Research Mentorship Award
- 2009 University of Pennsylvania Provost's Undergraduate Research Mentoring Program
- 2008 Dr. Yoram J. Kaufman Scholarship
- Maryland Distinguished Scholar Finalist
- National AP Scholar with Distinction

PEER-REVIEWED PUBLICATIONS

4. Huang, E.J., Fang, E.X., Hanley, D.F., and Rosenblum, M. (2017). Inequality in Treatment Benefits: Can We Determine if a New Treatment Benefits the Many or the Few? *Biostatistics* 18(2): 308-324. http://goo.gl/zcR4mM

3. Hanley, D. F., Thompson, R. E., Muschelli, J., Rosenblum, M., McBee, N., Lane, K., Bistran-Hall, A. J., Mayo, S. W., Keyl, P., Gandhi, D., Morgan, T. C., Ullman, N., Mould, W. A., Carhuapoma, J. R., Kase, C., Ziai, W., Thompson, C. B., Yenokyan, G., Huang, E., Broaddus, W. C., Graham, R. S., Aldrich, E. F., Dodd, R., Wijman, C., Caron, J-L., Huang, J., Camarata, P., Mendelow, A.D., Gregson, B., Janis, S., Vespa, P., Martin, N., Awad, I., Zuccarello, M., for the MISTIE II Investigators. (2016). Safety and efficacy of minimally invasive surgery plus alteplase in intracerebral haemorrhage evacuation (MISTIE): a randomized, controlled, open-label, phase 2 trial. *Lancet Neurology* 15(12):1228-1237. http://goo.gl/l0OWVP

   Role: I performed the statistical analysis of the treatment effect on the volume of clot removed.

2. Huang, E.J., Bonafide, C.P., Keren, R., Nadkarni, V.M., and Holmes, J.H. (2013). Medications associated with clinical deterioration in hospitalized children. *Journal of Hospital Medicine* 8(5): 254–260. http://goo.gl/bk9Usx

1. Bonafide, C.P., Roberts, K.E., Priestley, M.A., Tibbetts, K.M., Huang, E., Nadkarni, V.M., and Keren, R. (2012). Development of a pragmatic measure for evaluating and optimizing rapid response systems. *Pediatrics* 129(4): e874 -81. http://goo.gl/hvZHkO

PAPERS IN PREPARATION

2. Huang, E.J., and Varadhan, R. Modeling of Frailty and Other Aging Phenotypes Assessed at Discrete Study Visits

1. Huang, E.J., Fang, E.X., Hanley, D.F., and Rosenblum, M. Constructing a Confidence Interval for the Proportion who Benefit from Treatment Using Randomized Trial Data

SKILLS

- Programming and Software: R, MATLAB, Stata, CPLEX, LaTeX, Lyx

ORAL PRESENTATIONS

11. Invited Talk at SRCOS Summer Research Conference 2017: "Inequality in Treatment Benefits: Can We Determine if a New Treatment Benefits the Many or the Few?", Jekyll Island, GA, 6/7/17

10. Johns Hopkins Center on Aging and Health Research-in-Progress Session: "Inequality in Treatment Benefits: Can We Determine if a New Treatment Benefits the Many or the Few?", Baltimore, MD, 5/15/17

9. ENAR Spring 2017 Meeting: "Constructing a Confidence Interval for the Fraction who Benefit, Using a Randomized Trial", Washington D.C., 3/15/17

8. National Institute of Allergy and Infectious Diseases (NIAID) Seminar: "Inequality in Treatment Benefits: Can We Determine if a New Treatment Benefits the Many or the Few?", Rockville, MD, 2/21/17

7. Johns Hopkins Biostatistics 622 Course Lecture (as a substitute for Prof. Bandeen-Roche) "Statistical Inference about Proportions", Baltimore, MD, 10/27/16

6. Johns Hopkins Causal Inference Seminar: "Inequality in Treatment Benefits: Can We Determine if a New Treatment Benefits the Many or the Few?", Baltimore, MD, 5/10/16

5. ENAR Spring 2016 Meeting: "Inequality in Treatment Benefits: Can We Determine if a New Treatment Benefits the Many or the Few?", Austin, TX, 3/7/16

4. Johns Hopkins Center on Aging and Health Research-in-Progress Session: "Modeling of Aging Phenotypes Assessed at Discrete Study Visits", Baltimore, MD, 2/8/16

3. Johns Hopkins Center on Aging and Health Research-in-Progress Session: "Inequality in Treatment Benefits: Can We Determine if a New Treatment Benefits the Many or the Few?", Baltimore, MD, 4/20/15

2. ENAR Spring 2015 Meeting: "Estimating the Fraction who Benefit from a Treatment, using Randomized Trial Data", Miami, FL, 3/16/15

1. Johns Hopkins Center on Aging and Health Research-in-Progress Session: "Prediction Modeling for Grouped Time-to-Event (Visit) Data", Baltimore, MD, 3/24/14

POSTER PRESENTATIONS

2. The Gerontological Society of America's 68th Annual Scientific Meeting: "Prediction Modeling of Aging Phenotypes Assessed at Discrete Study Visits", Orlando, FL, 11/18/15

1. Pediatric Hospital Medicine Conference: "Medication sentinels associated with clinical deterioration in hospitalized children", Kansas City, MO, 7/29/11

TEACHING

Johns Hopkins University

- 2014, 2015, 2016 Lead Teaching Assistant for Statistical Methods in Public Health I: teach two 1.5-hour labs per week, proctor and help write exams, hold office hours, serve as substitute lecturer for Prof. Bandeen-Roche

  Student reviews:

  - "She went at a great pace, very thoroughly described the logic behind all of the concepts, and answered all questions directly."
  - "I've found your labs/lectures the clearest of anyone's."

- 2014, 2015, 2016 Teaching Assistant for Statistical Methods in Public Health II: hold office hours, grade homework and exams

- 2014, 2016, 2017 Teaching Assistant for Statistical Methods in Public Health III: hold office hours, grade homework and exams
- 2014, 2015 Teaching Assistant for Statistical Methods in Public Health IV: provide consulting services to graduate students on a data analysis project
- 2013 Teaching Assistant for Biostatistics in Public Health (undergraduate course): teach one 1-hour recitation per week, grade homework and exams

University of Pennsylvania

- 2011 Undergraduate Teaching Assistant for Bioengineering Signals and Systems: grade homework and exams

LEADERSHIP AND SERVICE

- 2015-2016 Co-chair of the Johns Hopkins Gerontology Interest Group
- 2014-2016 Organizer of Johns Hopkins Department of Biostatistics TA Training Day
- 2014 Volunteer for ENAR Conference in Baltimore, MD

MEMBERSHIP

- 2015-present Eastern North American Region (ENAR)
- 2014-present American Statistical Association
- 2014-present Institute of Mathematical Statistics
- 2014-present Johns Hopkins Department of Biostatistics Causal Inference Working Group